U.S. ARMY RDECOM®

UNCLASSIFIED: Distribution Statement A.
Approved for public release

TARDEC

TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Computational Fluids Domain Reduction to a Simplified Fluid Network

19 April 2012

Project Lead:
Rob E Smith
rob.e.smith@us.army.mil
586-282-4121

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **09 JUL 2012**	2. REPORT TYPE **Briefing Charts**	3. DATES COVERED **09-07-2012 to 09-07-2012**

4. TITLE AND SUBTITLE **Computational Fluids Domain Reduction to a Simplified Fluid Network**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) **Robert Smith**	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army TARDEC ,6501 E.11 Mile Rd,Warren,MI,48397-5000**	8. PERFORMING ORGANIZATION REPORT NUMBER **#22880**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) **U.S. Army TARDEC, 6501 E.11 Mile Rd, Warren, MI, 48397-5000**	10. SPONSOR/MONITOR’S ACRONYM(S) **TARDEC**
	11. SPONSOR/MONITOR’S REPORT NUMBER(S) **#22880**

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
briefing charts

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Public Release**	**81**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Agenda

- About Rob E Smith
- Research Objectives
- Quick Introduction to Data Mining
- cfdMine Tool
- cfdMine Mahalanobis
- Validation Results
- Conclusions/ Next Steps

About Me

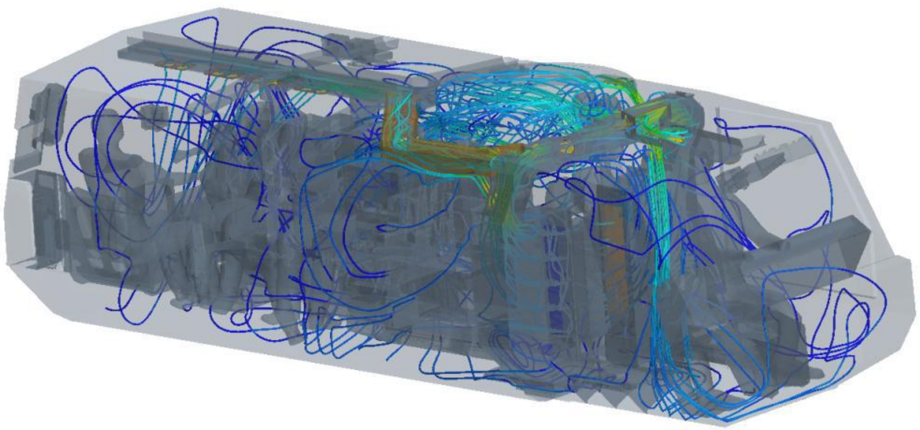
- Career:
  - Present: Team Lead for the CFD/Signature Modeling,TARDEC
  - 2003-2008 Survivability Specialist, General Dynamics Land Systems
  - 1999-2003 Research Engineer, ThermoAnalytics, Inc.
  - 1994-1996 Test Engineer/ Lab Tech, Whirlpool Corporation
- Education:
  - 2002 MSME Michigan Tech
    - Project: Infrared Signature Modeling of the AH-64 Apache
  - 1999 BSME Michigan Tech
- Classes Taken Specific to This Thesis:
  - Data Mining
  - Neural Networks
  - Intelligent Systems

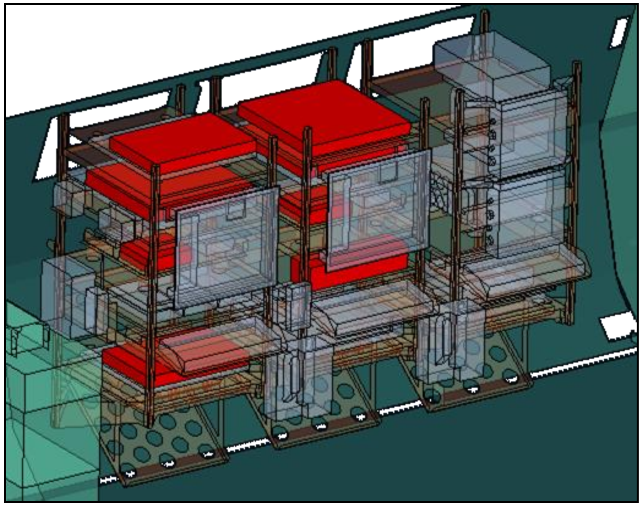
Agenda

- About Rob E Smith
- Research Objectives
- Quick Introduction to Data Mining
- cfdMine Tool
- cfdMine Mahalanobis
- Validation Results
- Conclusions/ Next Steps

Basic Problem

Typical Military Vehicle HVAC Simulation



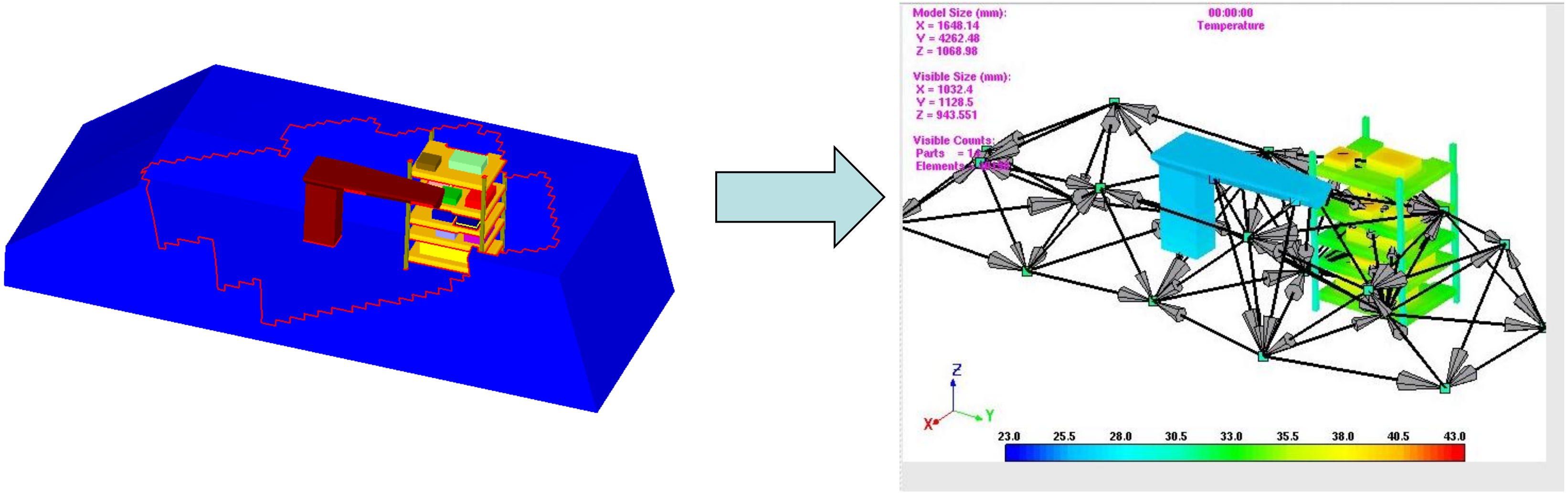


Typical Military Vehicle Equipment Rack

- CFD models of vehicle interior are ~15 million cells
- Solve steady state HVAC with 64 cores overnight
- Need to perform heat soaked vehicle simulation and HVAC startup
  - Would take at least a week
- Particularly interested in equipment
  - Want operational temps in less than 15 minutes

Research Objectives

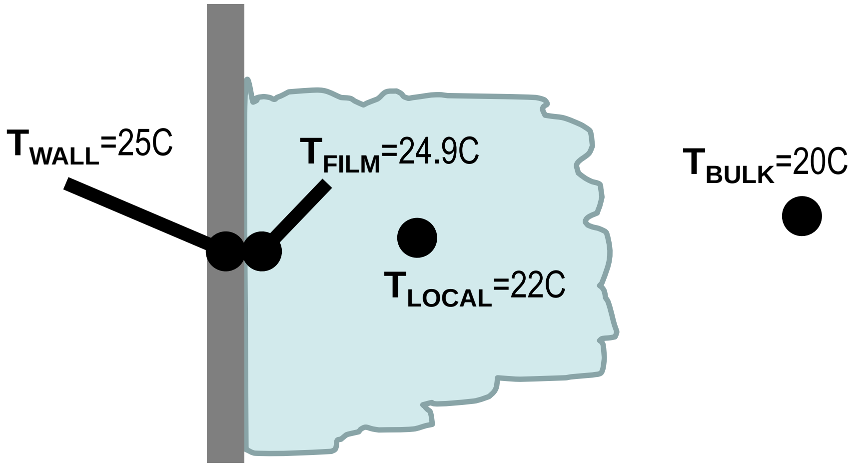
1. Demonstrate process of clustering a <u>steady-state</u> CFD domain into a <u>transient</u> lumped fluid network simulation



Research Objectives

1. Demonstrate process of clustering a steady-state CFD domain into a transient lumped fluid network simulation
2. Show value of remapping CFD fluid film convection to localized fluid nodes

Which H?



Remapping Equation

$$h_{film}(T_{film} - T_{wall}) = \dot{q} = h_{local}(T_{local} - T_{wall})$$

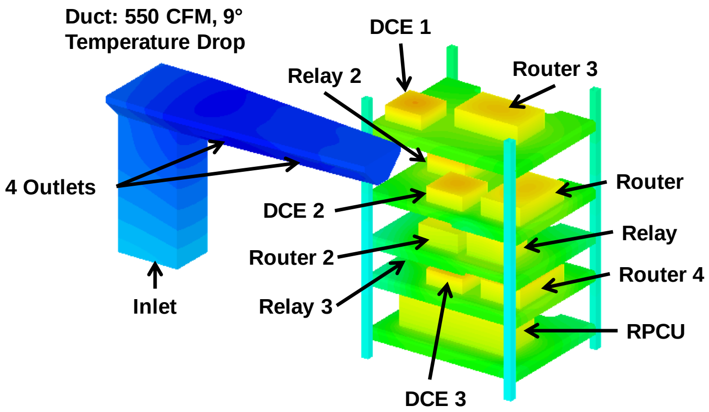
so

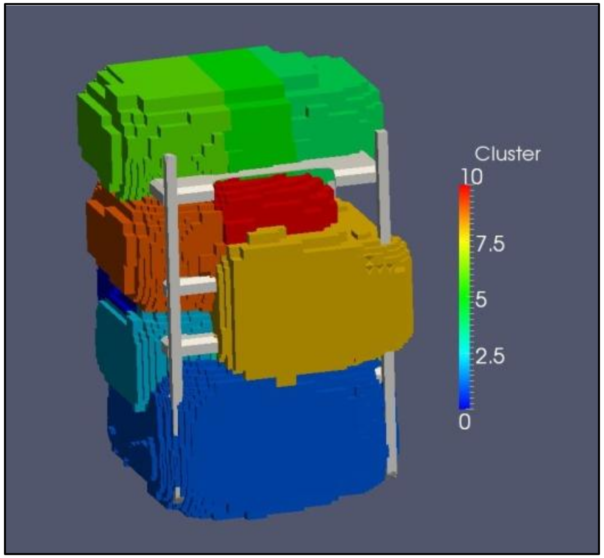
$$h_{local} = \frac{h_{film}(T_{film} - T_{wall})}{T_{local} - T_{wall}}$$

TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Research Objectives

1. Demonstrate process of clustering a steady-state CFD domain into a transient lumped fluid network simulation
2. Show value of remapping CFD fluid film convection to localized fluid nodes
3. Investigate the use of clustering to track temperatures around specific equipment or locations

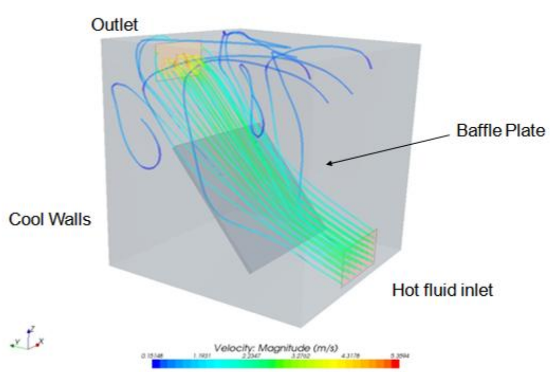


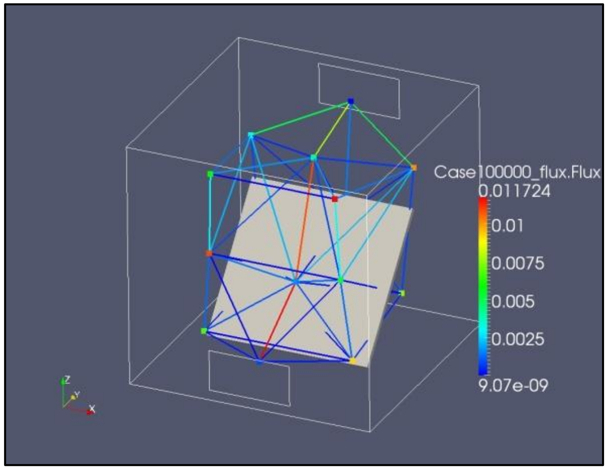


TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Research Objectives

1. Demonstrate process of clustering a <u>steady-state</u> CFD domain into a <u>transient</u> lumped fluid network simulation
2. Show value of remapping CFD fluid film convection to localized fluid nodes
3. Investigate the use of clustering to track temperatures around specific equipment or locations
4. Show new ways of flowfield visualization using clustering

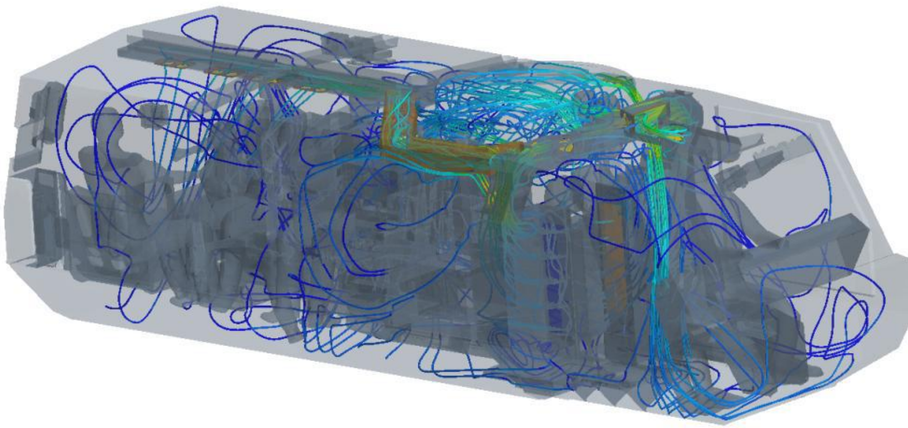


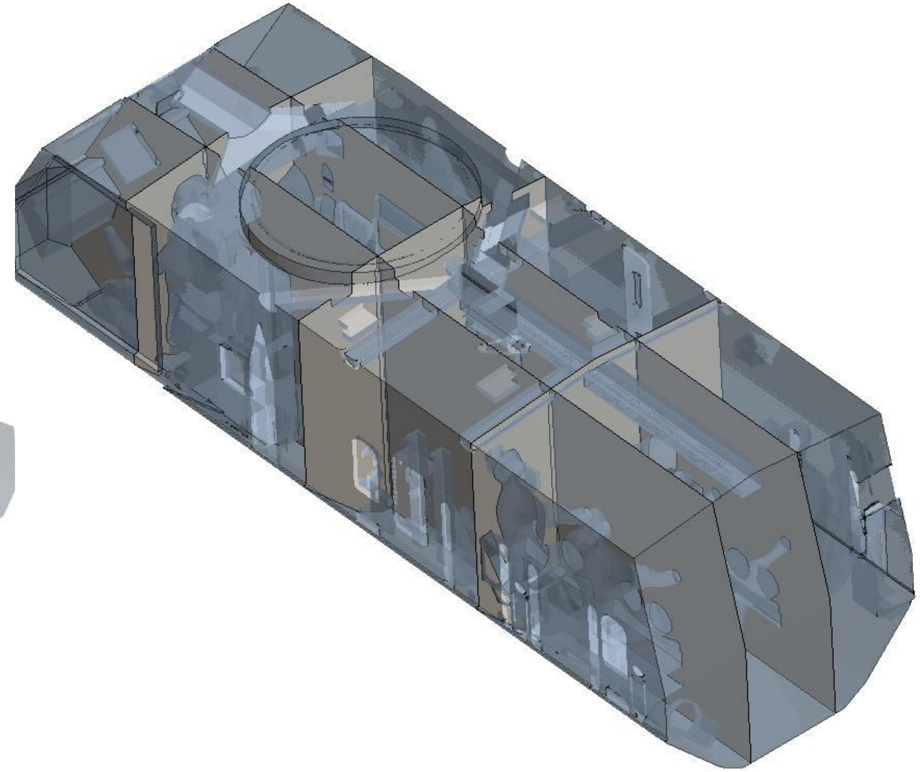


Research Objectives

1. Demonstrate process of clustering a steady-state CFD domain into a transient lumped fluid network simulation
2. Show value of remapping CFD fluid film convection to localized fluid nodes
3. Investigate the use of clustering to track temperatures around specific equipment or locations
4. Show new ways of flowfield visualization using clustering
5. Validate subvoluming and new clustering techniques against fully transient CFD data

State of the Art: Subvoluming Approach

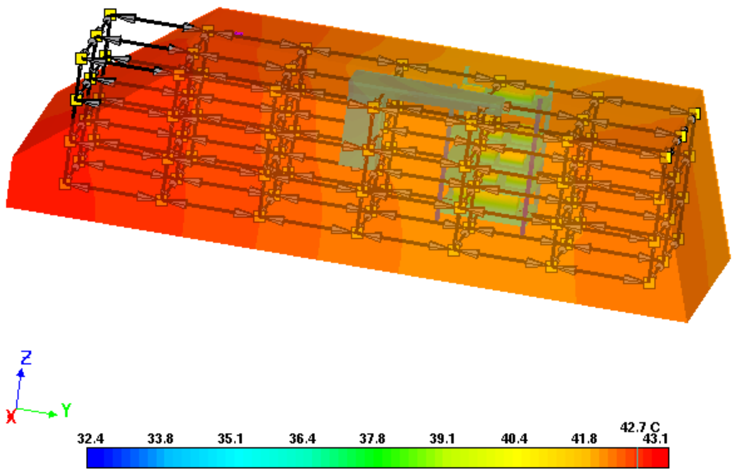




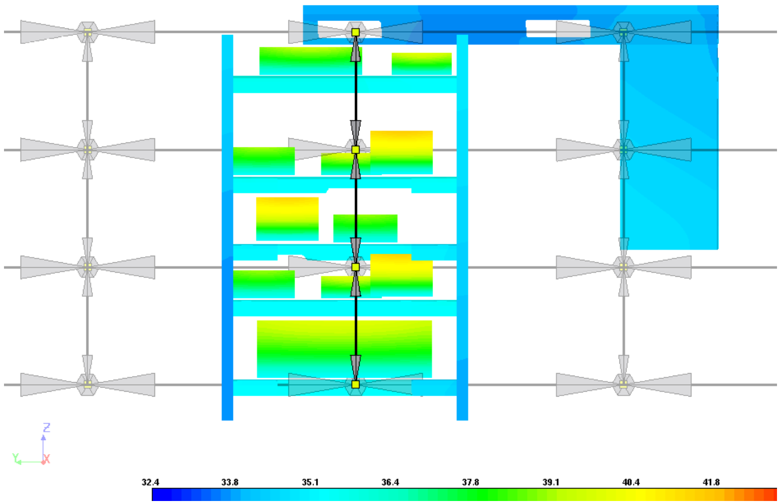
- Simple idea: cut planes through the domain and track volume fluxes across surfaces to make control volume network.
- Apply a fixed 15 w/m^2 convection coefficient that convects to closest node

Subvolumed Model / Problems

Subvolume nodal arrangement
72 Nodes Evenly Spaced

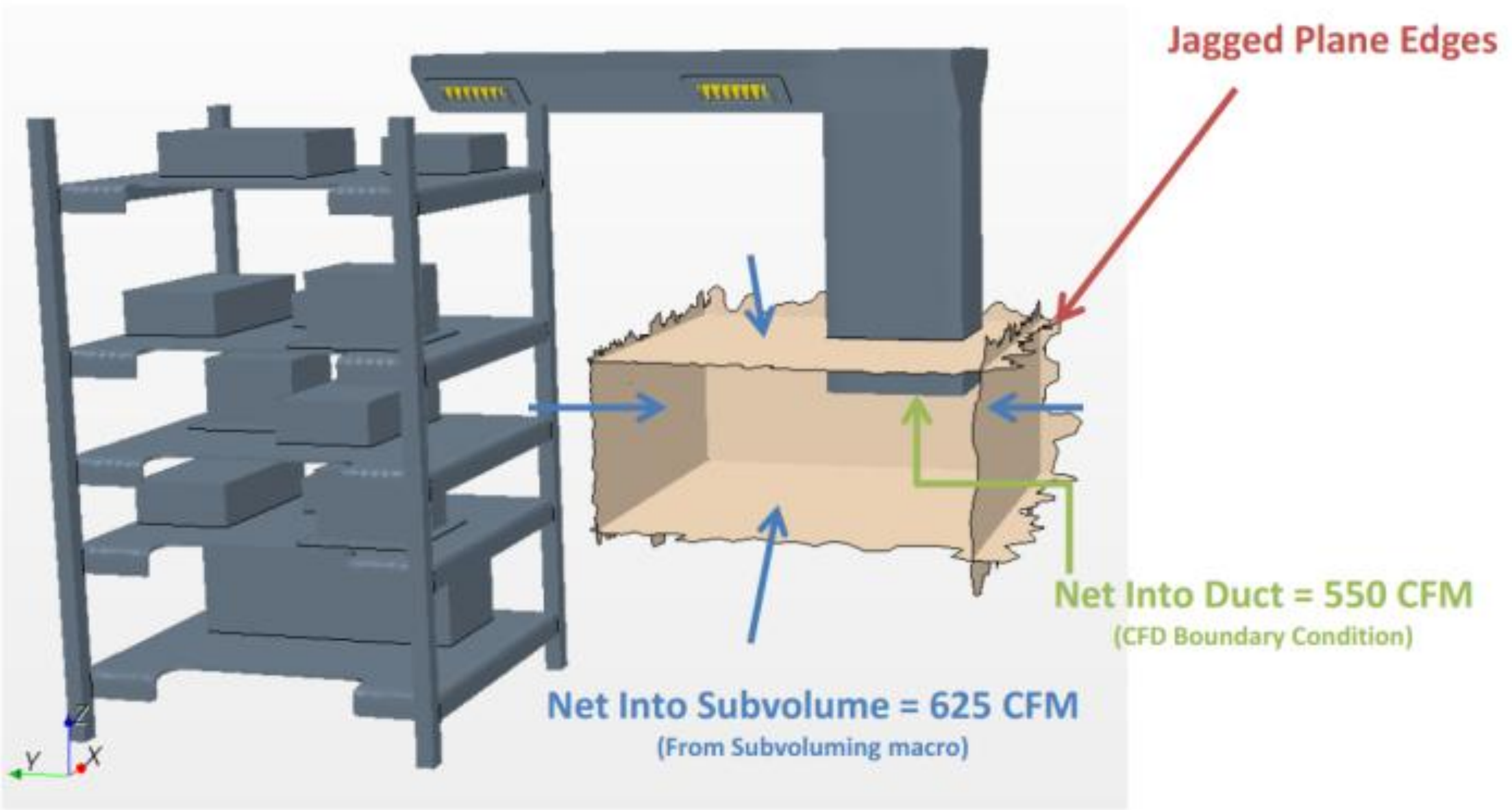


Equipment Rack



- Volumes cut complete across the shelves.
- Multiple equipment associated with each node

Another Problem with Subvoluming



Agenda

- About Rob E Smith
- Research Objectives
- Quick Introduction to Data Mining
- cfdMine Tool
- cfdMine Mahalanobis
- Validation Results
- Conclusions/ Next Steps

Data Mining

- Data mining – knowledge discovery in large sets of data
- Relatively new field in the area of computer science
- Uses/ Examples:
  - Grouping similar documents together
  - Determining consumer spending patterns (market segmentation)
  - Risk analysis and management
  - Fraud detection and detection of unusual patterns (outliers)
  - DNA and bio-data analysis

Information Harvesting

Knowledge Mining

Data Mining

Knowledge Discovery in Databases

Data Dredging

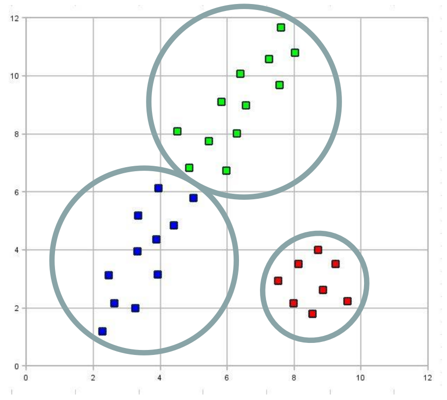
Data Pattern Processing

Knowledge Extraction

Database Mining

Data Mining - Clustering

- Clustering is type of data mining
- Unsupervised classification of patterns observations data items or feature vectors into groups or clusters - Jain [9]
- Clusters should be externally isolated and internally cohesive, implying homogeneity within clusters and heterogeneity between clusters - Cormack [10]
- Examples of use
  - pattern recognition
  - image processing
  - market research
  - document classification for web searching



Factors Involving Data Mining

- Selection of features of interest
  - Quantitative features:
    - Continuous values (e.g. weight).
    - Discrete values (e.g. number of computers)
    - Interval values (e.g. the duration of an event)
  - Qualitative features:
    - Nominal or unordered (e.g. color)
    - Ordinal (e.g. qualitative evaluations of temperature "cool" or "hot")
- Preprocessing method
  - Z-normalization
  - Outlier rejection
- Selection of objective (goal) function
  - Maximize separation
  - Minimize within cluster similarity

Distance Metrics

Distances are normally used to measure similarity or dissimilarity between two data objects

- Equation 1: Euclidean Distance

$$d_e = \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2 \right)^{\frac{1}{2}}$$

- Equation 2: Minkowski Distance

$$d_m = \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^p \right)^{\frac{1}{p}}$$

- Equation 3: Cosine Similarity (dot product)

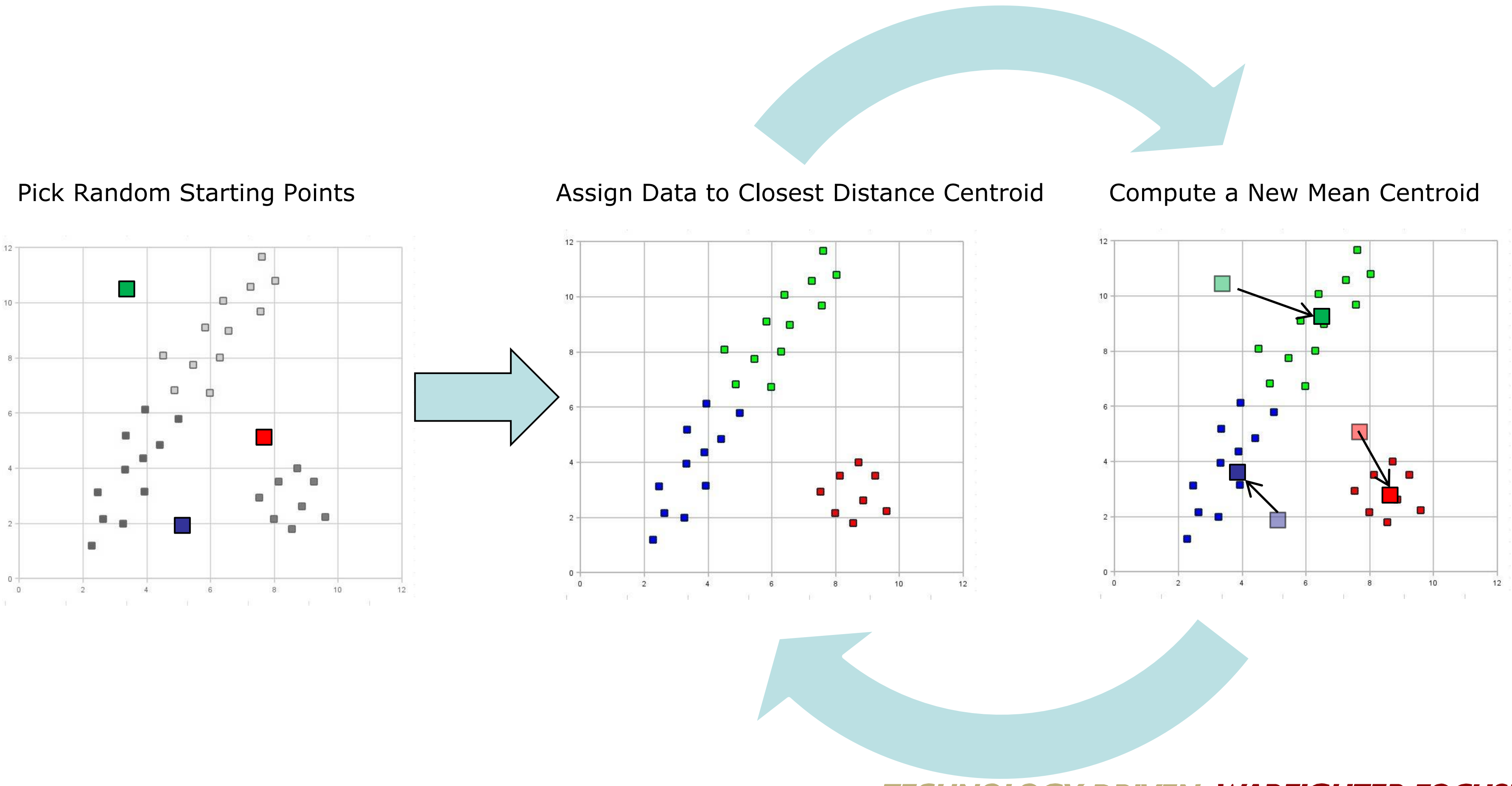
$$d_e = \left(\sum_{k=1}^{d} x_{i,k} \bullet x_{j,k}) \right)$$

- Equation 4: Manhattan Distance

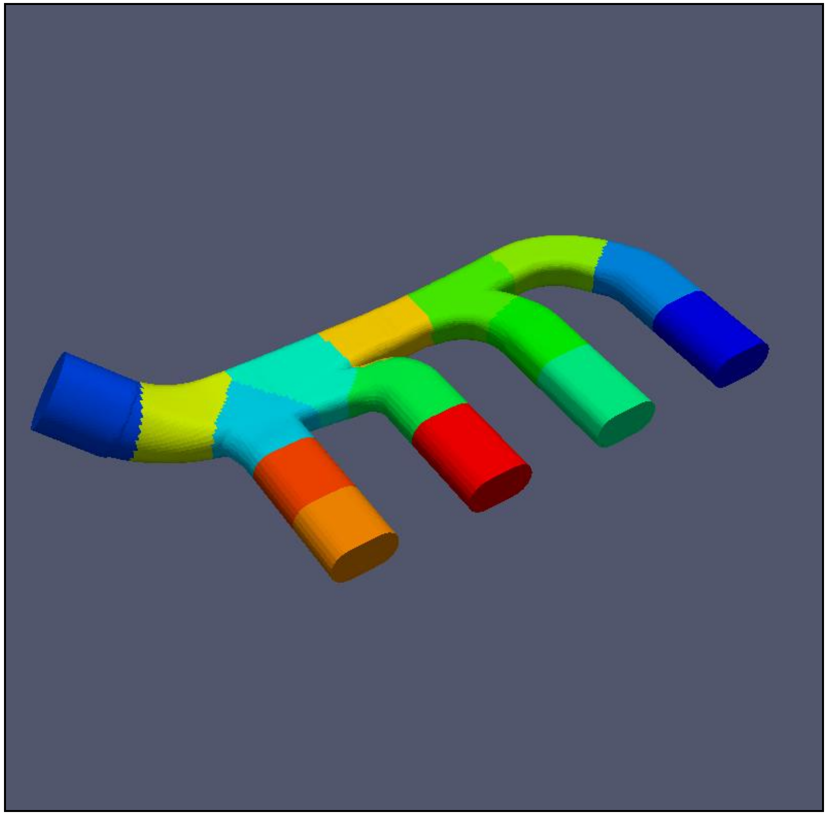
$$d_m = \sum_{k=1}^{d} (x_{i,k} - x_{j,k})$$

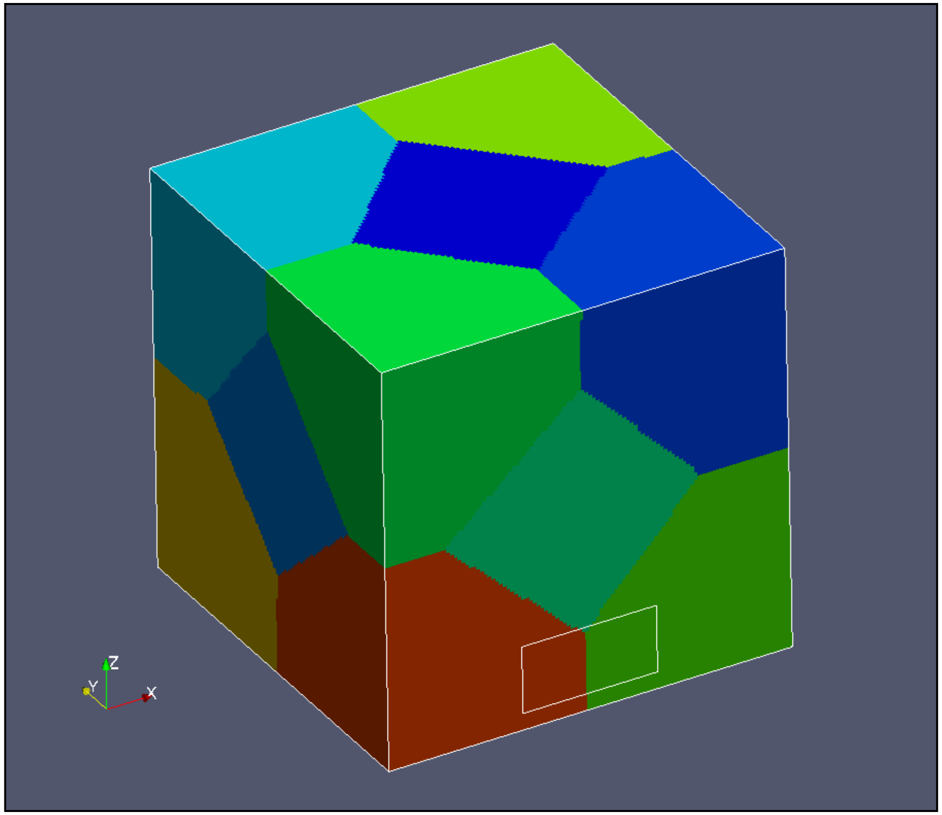
K-Means Basic Algorithm

Objective is to minimize distance to cluster centers based on a Euclidean distance.



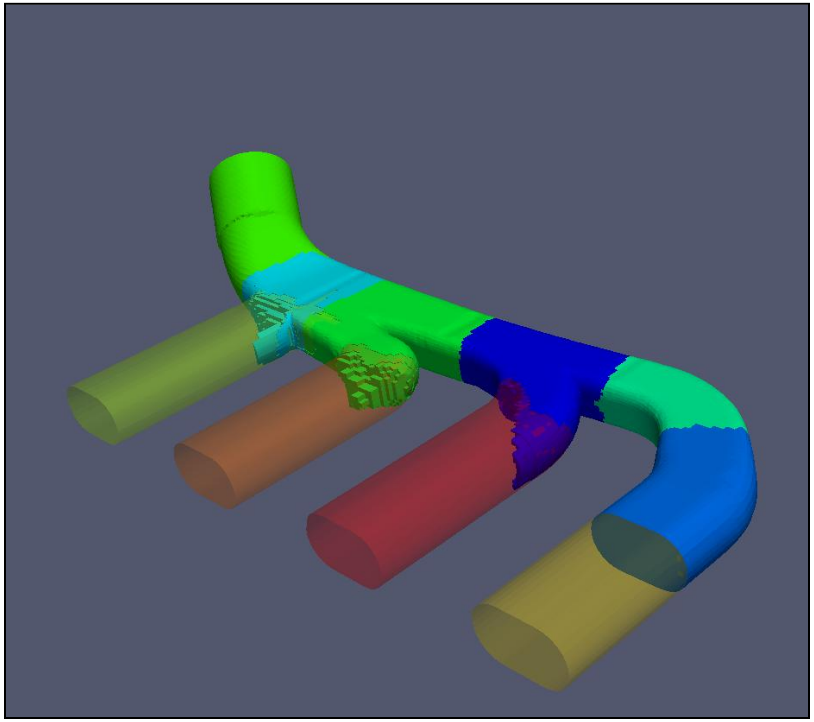
K-Means With no Feature Weighting (Only x,y,z)

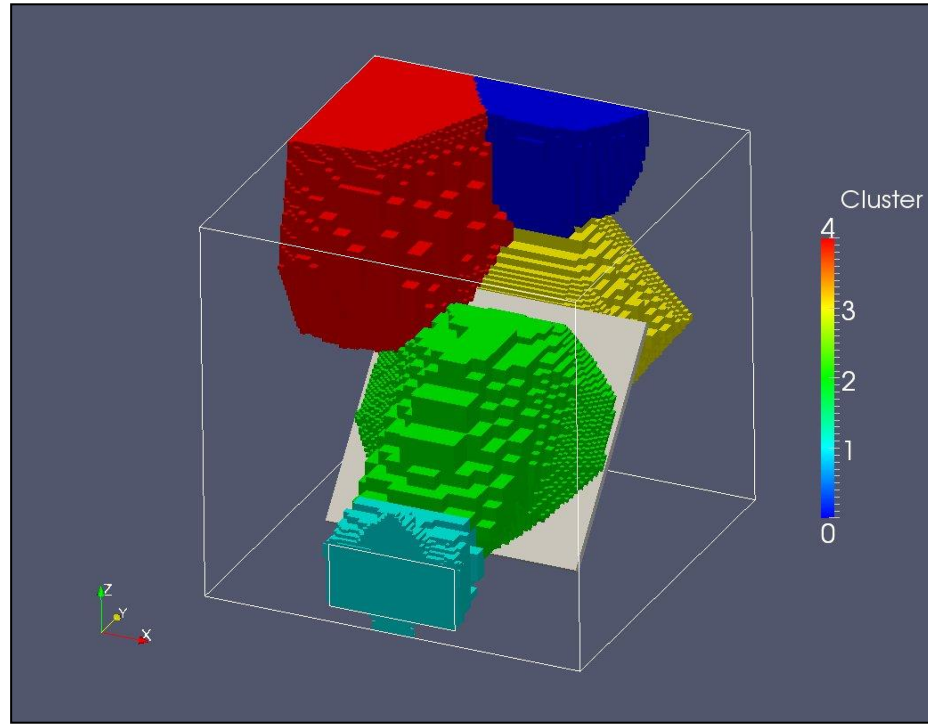




Purely based on x,y,z as features, k-means produces equal "inertia" clusters about each centroid

K-Means With Weighting and Multiple Variables



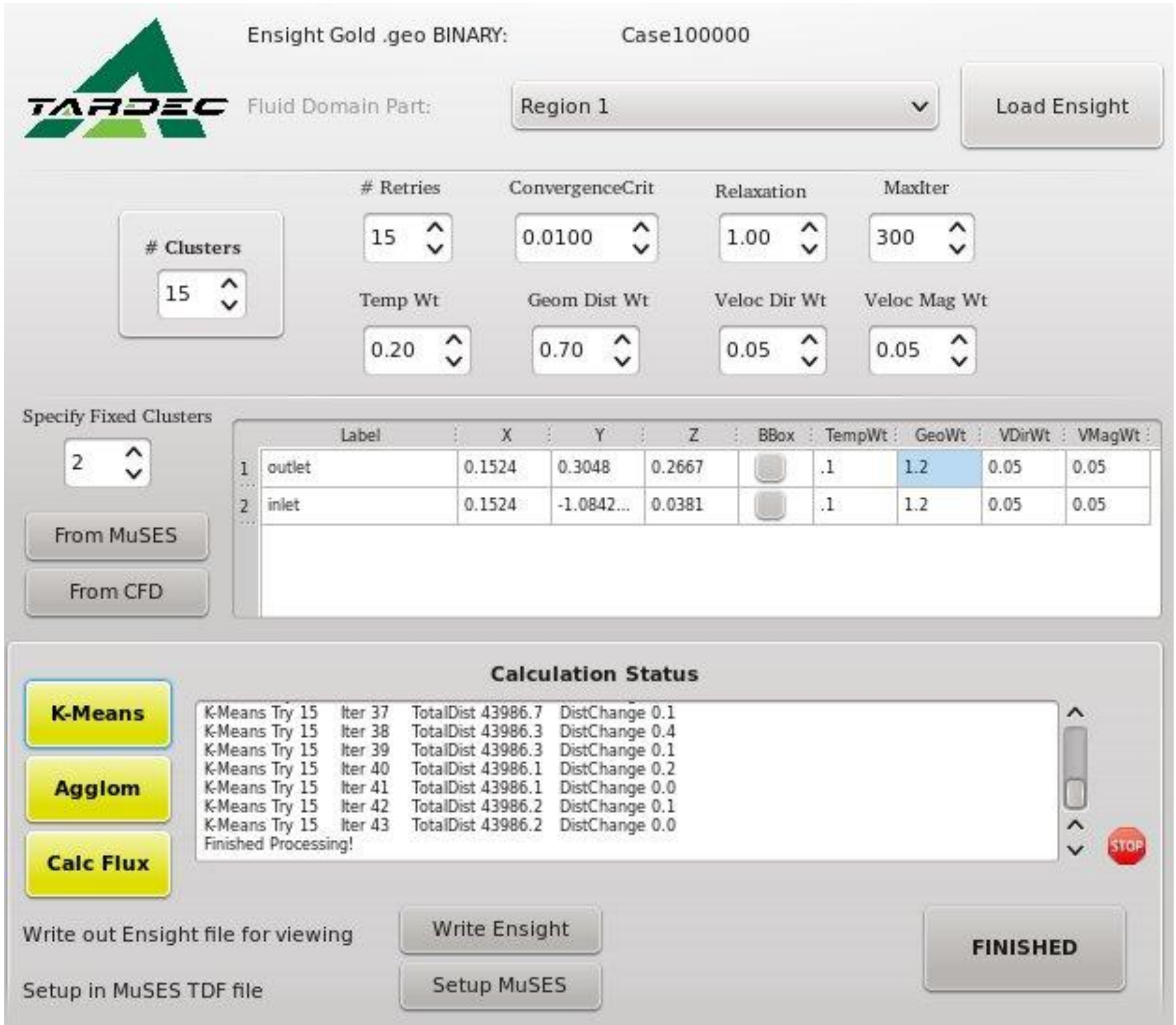


- Produces equal "inertia" clusters about the weighted multivariate feature dimensions.
- Weight x,y,z distance much higher to keep clusters compact.

Agenda

- About Rob E Smith
- Research Objectives
- Quick Introduction to Data Mining
- cfdMine Tool
- cfdMine Mahalanobis
- Validation Results
- Conclusions/ Next Steps

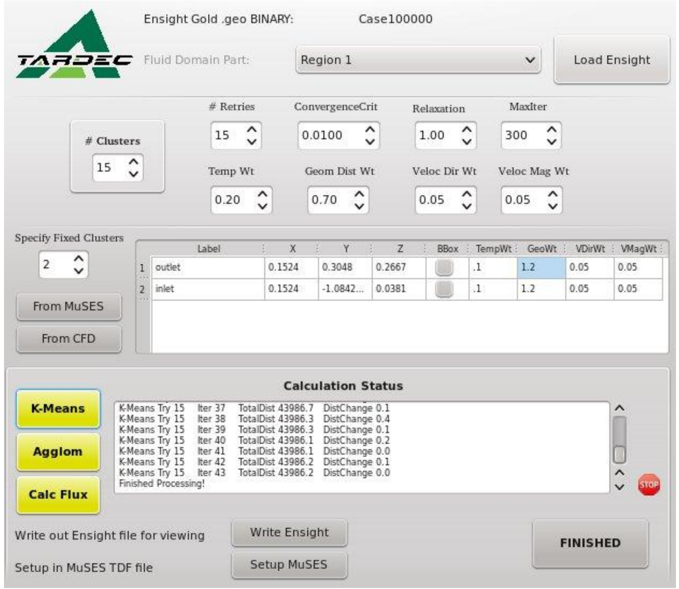
cfdMine GUI



TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

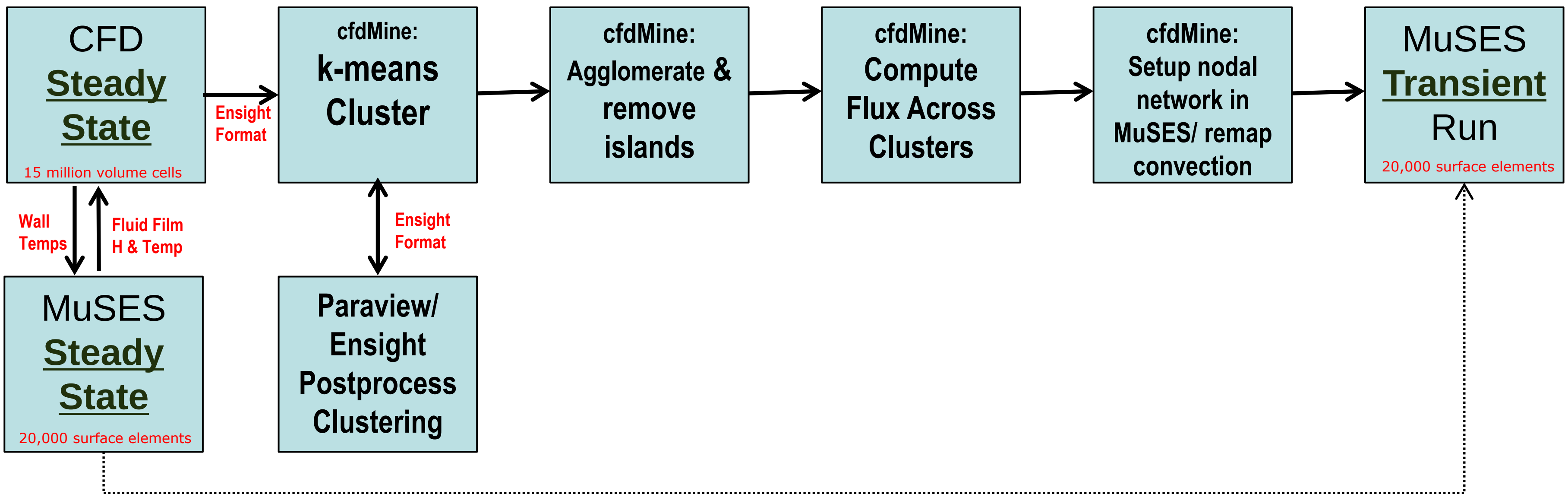
cfdMine Design

- Commercial-quality C++ tool
  - QT for user interface
  - Eigen stl library for Matrices
  - Kd-tree for matching elements
  - Modified dated Ensight reader for polyhedral mesh
  - 5000+ lines of code
- K-Means Implemented with Euclidean Distance
- Weighted features of interest
  - Temperature
  - X,Y,Z distances
  - Velocity magnitude
  - Velocity direction (as dot product)
- Preprocessing method
  - Z-normalization used



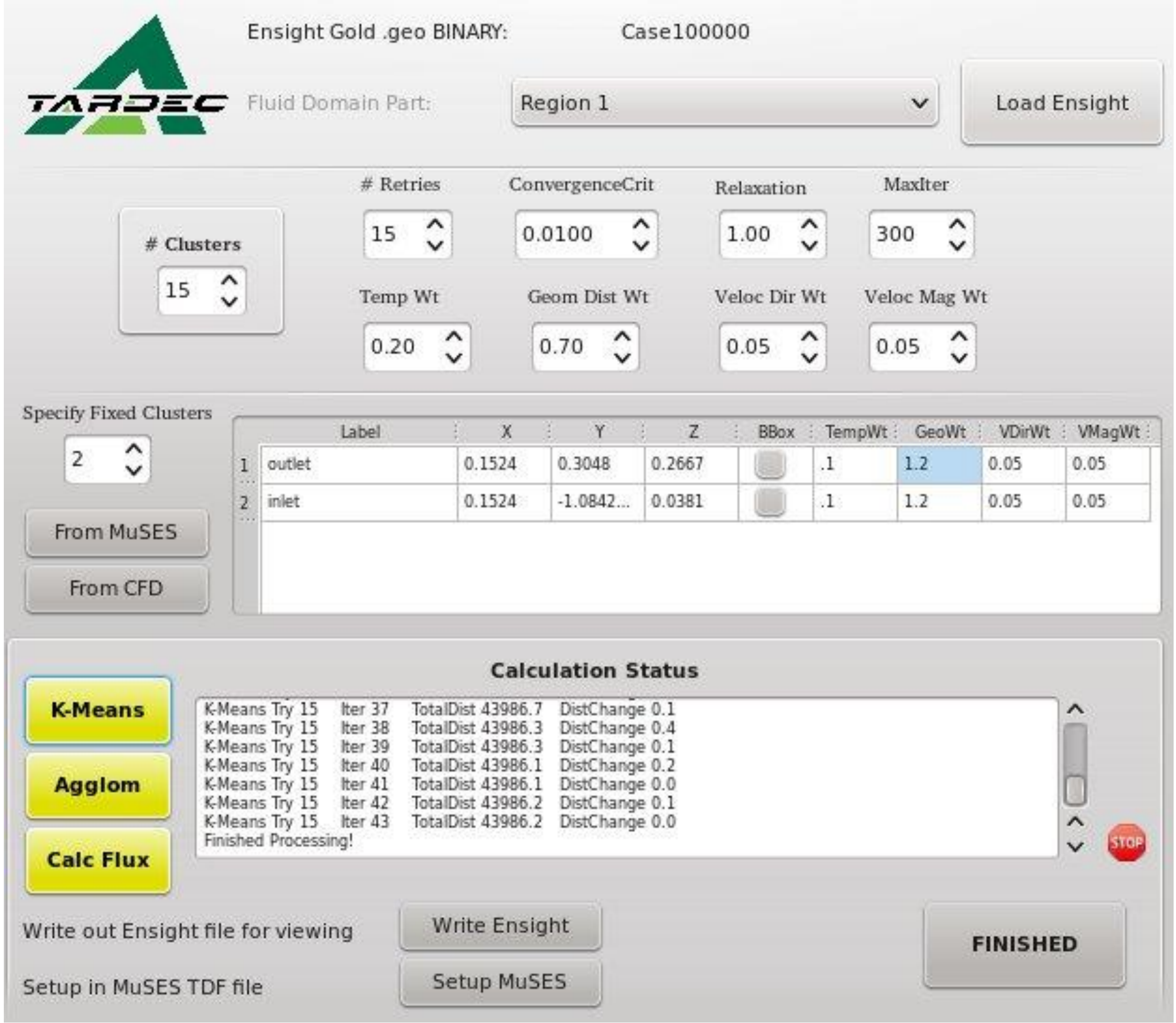
TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Basic cfdMine Process

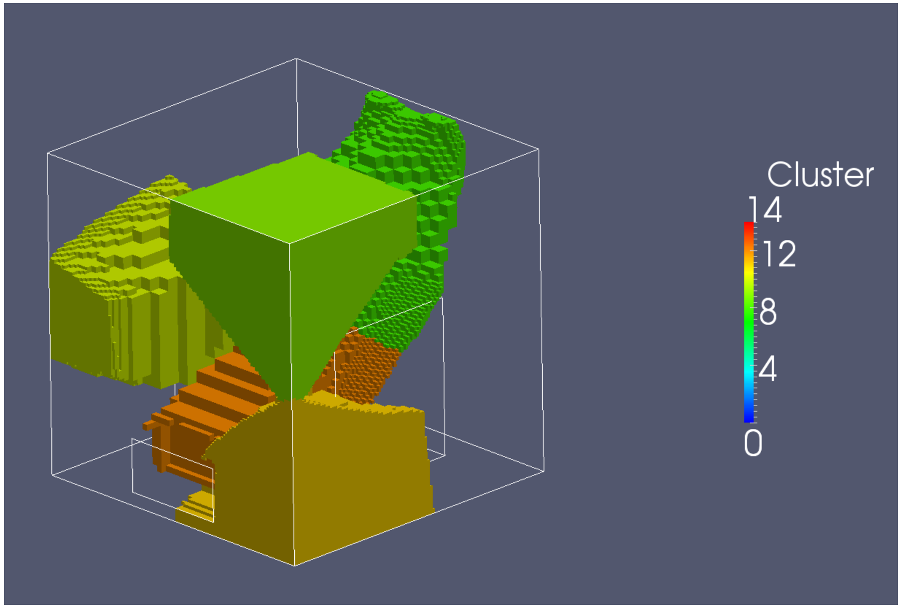


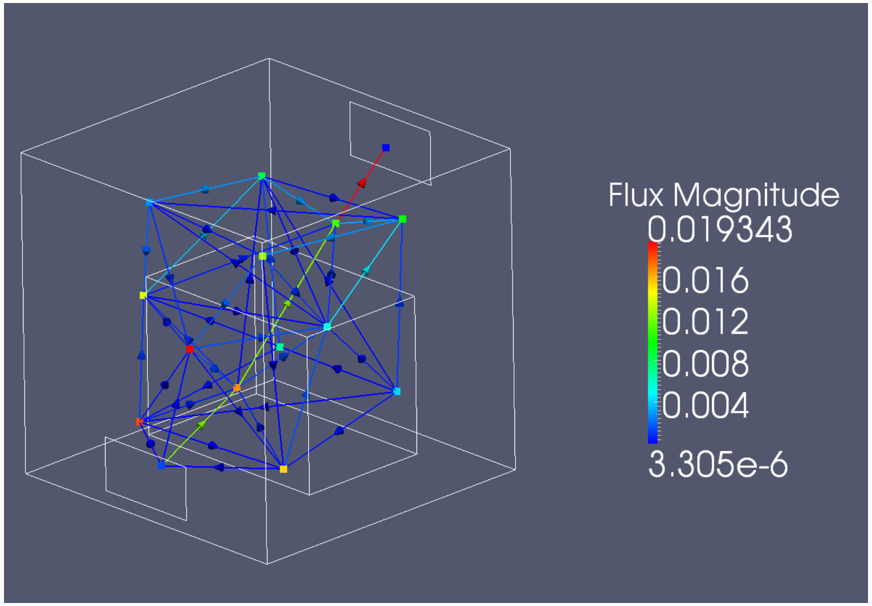
Note: The nodal network is set up in MuSES using a thermal link from every element to each cluster fluid node. (A first, but not ideal)

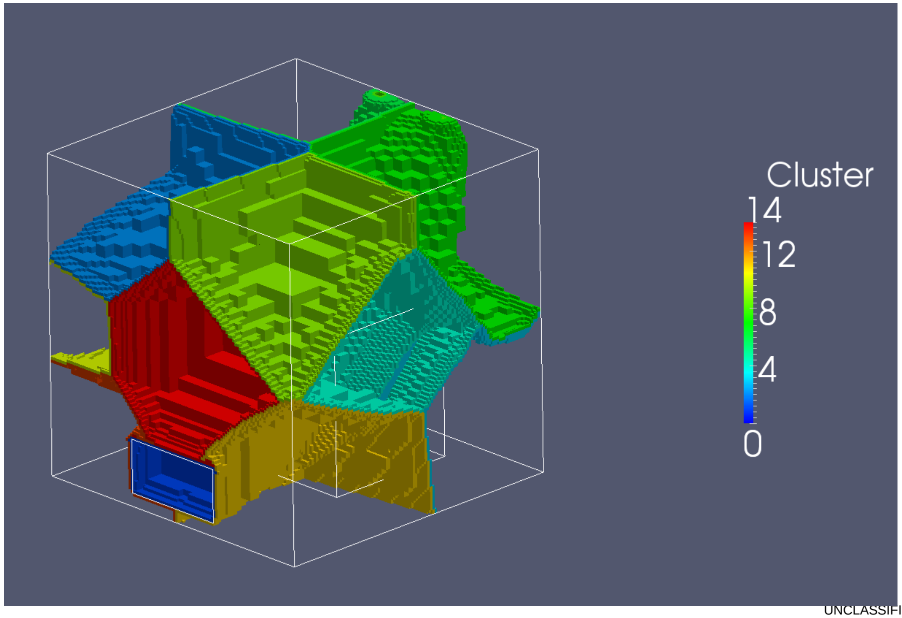
Which Knob Does What



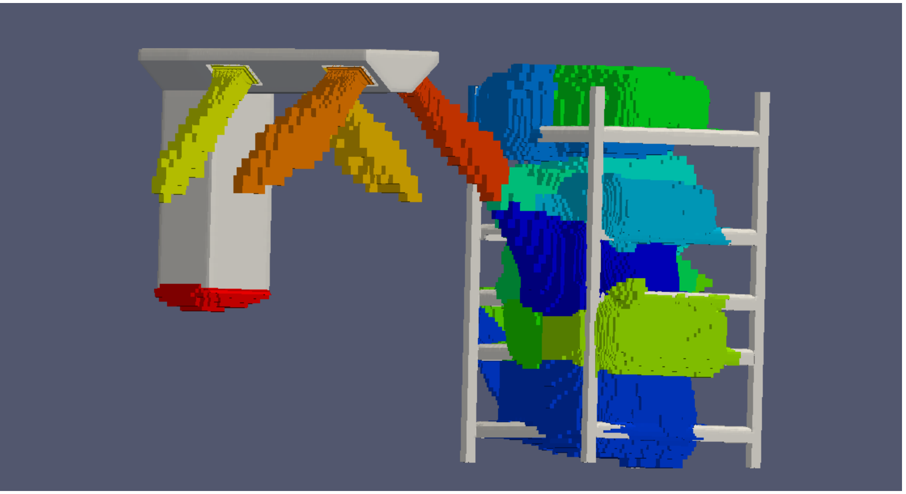
Visualizing Clusters – Case 1

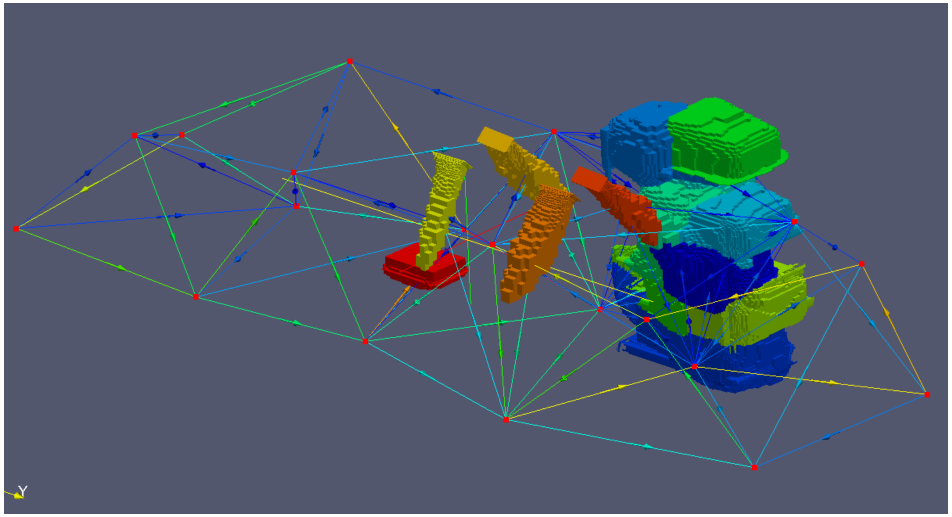


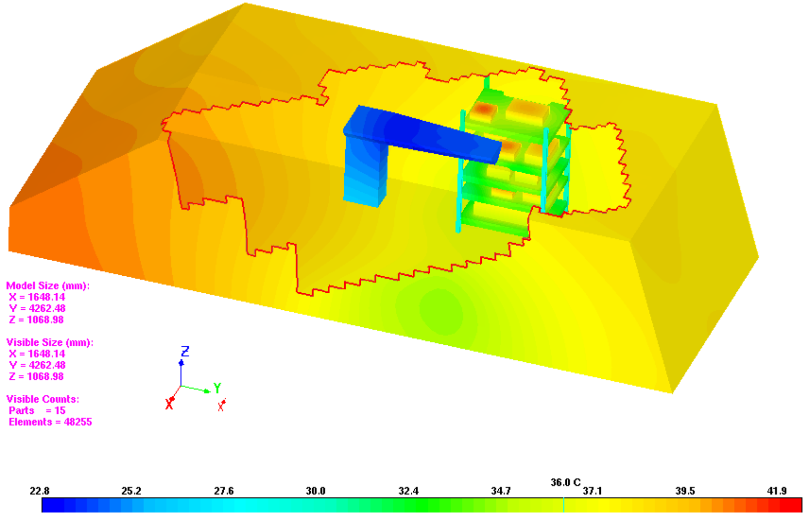


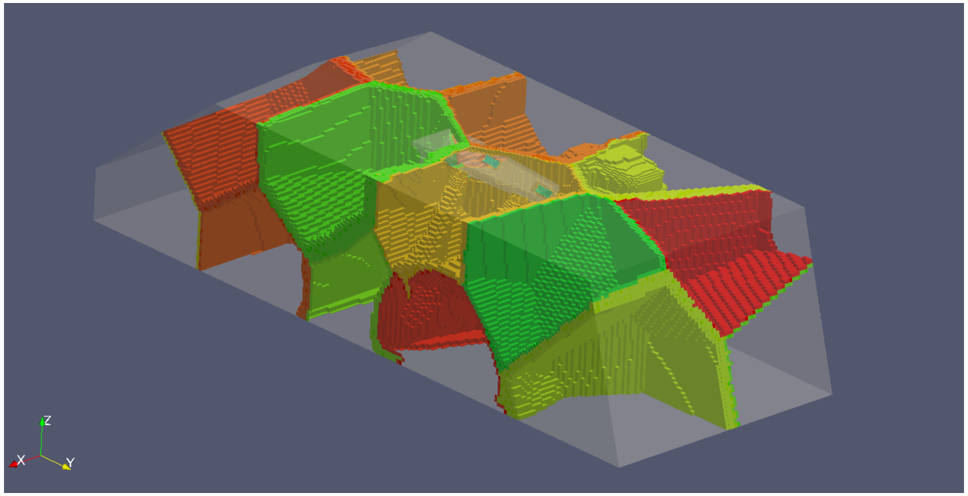


Visualizing Clusters – Generic Vehicle Interior

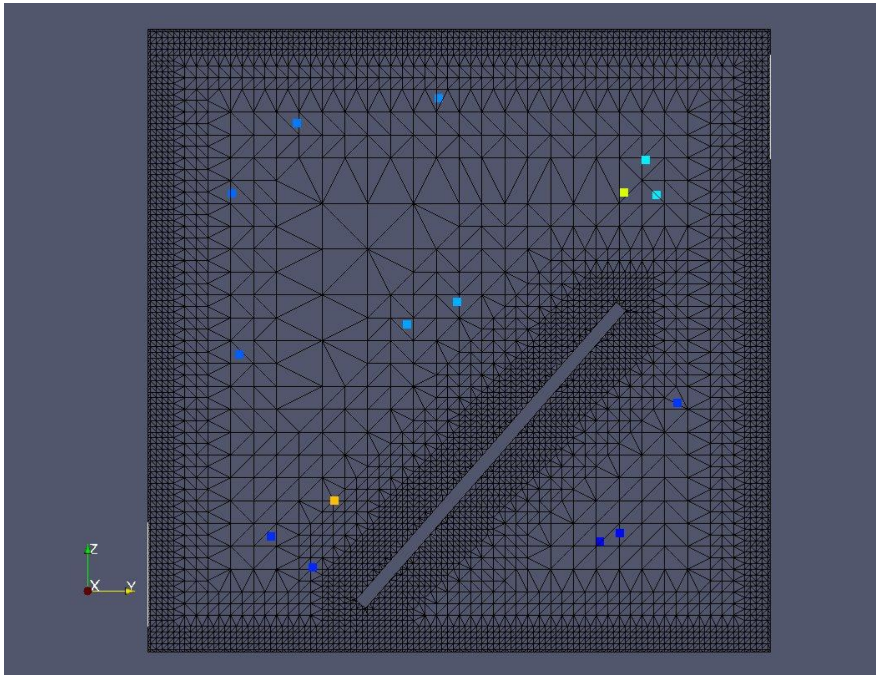


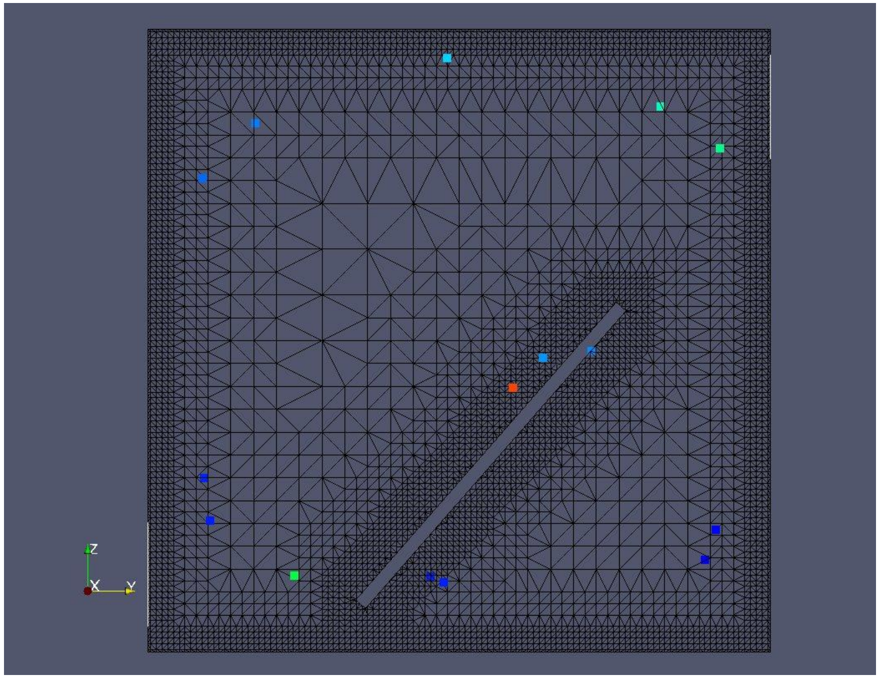






Importance of Volume Weighting

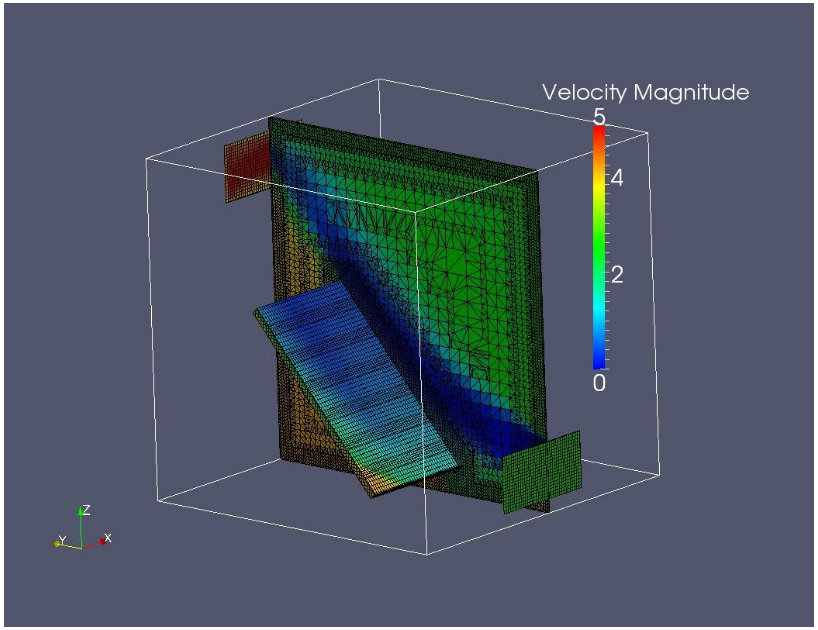


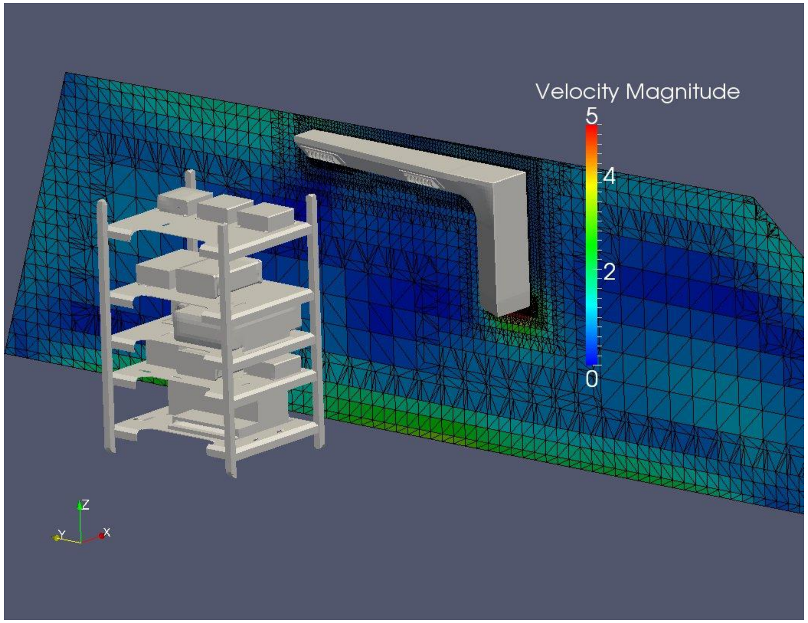


K-Means Volume Weighted Iterative Function

$$\bar{m}_l = \frac{\sum_{x \in S_l} v_x}{n_l} \sum_{x \in S_l} \vec{x} \cdot v_x$$

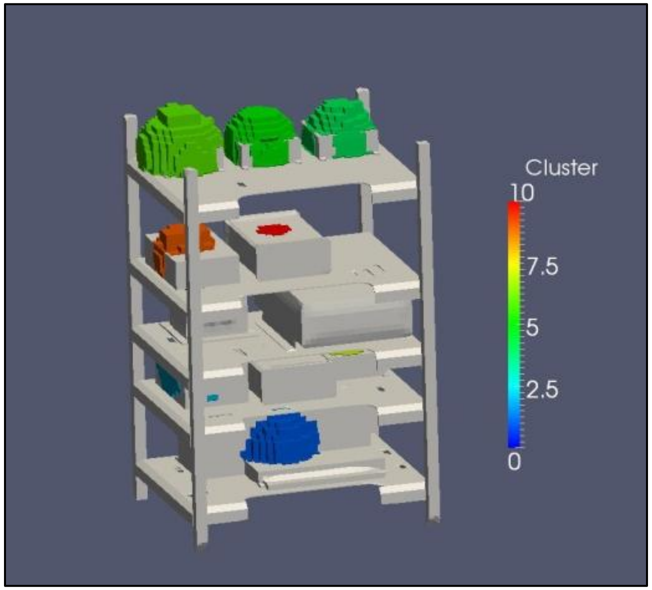
Reason To Weight – Typical Meshes



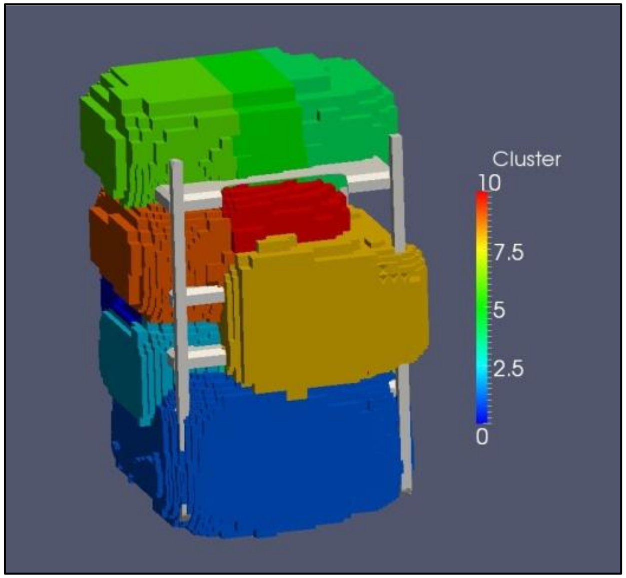


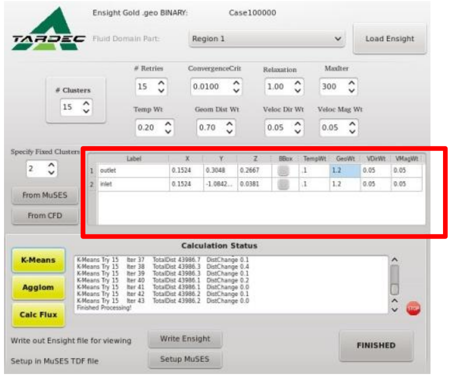
Fixing Clusters Around Equipment

Clustering About the Centroid



Clustering About the Bounding Box



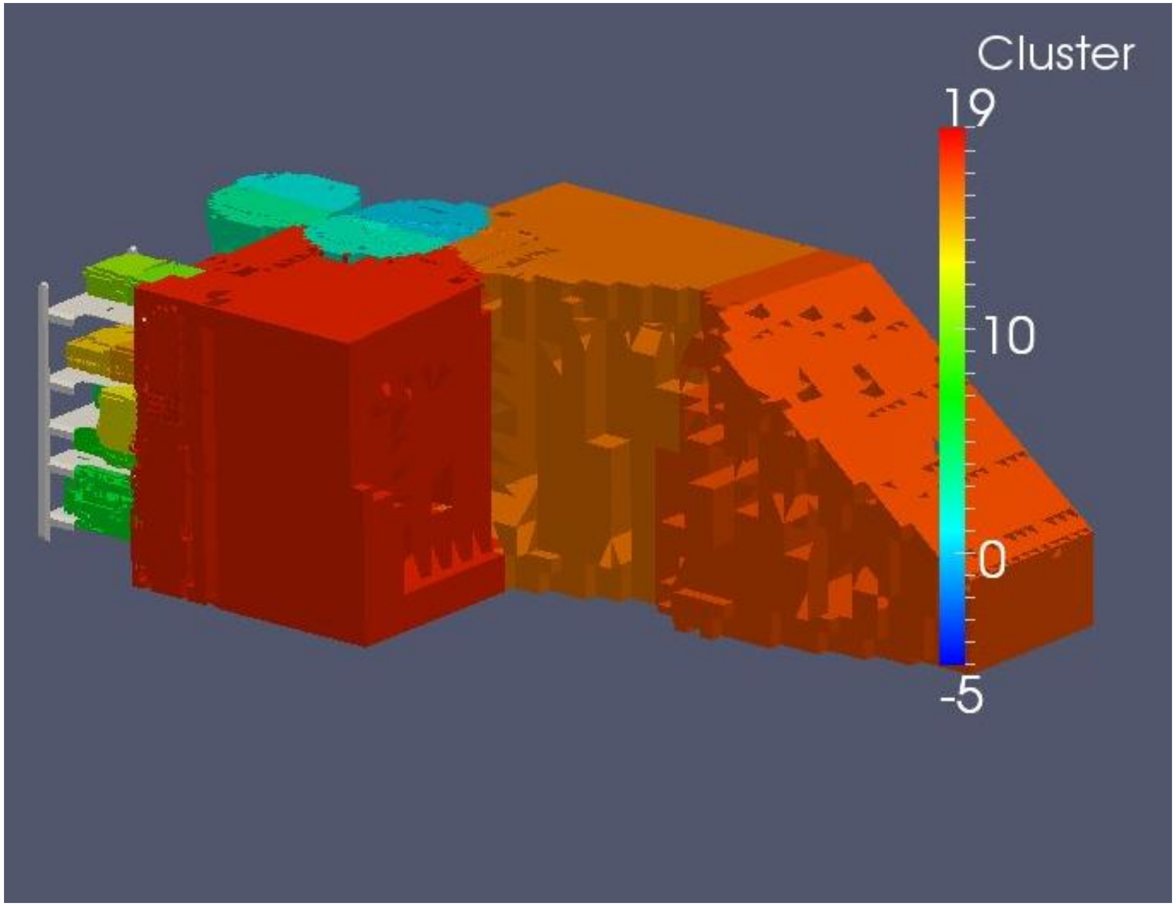


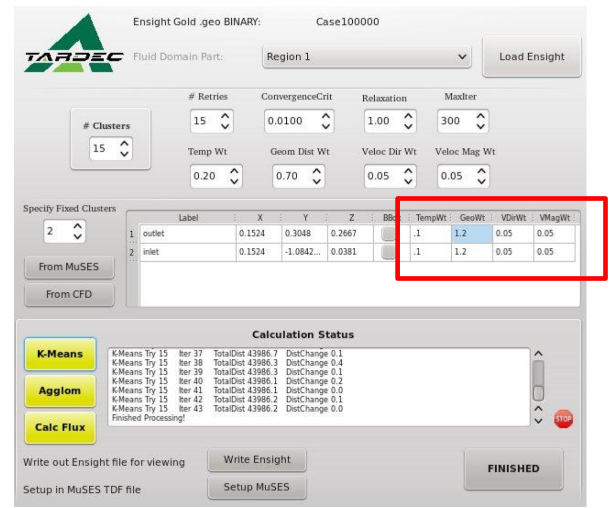
$$x_{bbox\_min} = (x < x_{min})\,?\,x_{min} : (x > x_{max})\,?\,x_{max} : x$$
$$y_{bbox\_min} = (y < y_{min})\,?\,y_{min} : (y > y_{max})\,?\,y_{max} : y$$
$$z_{bbox\_min} = (z < z_{min})\,?\,z_{min} : (z > z_{max})\,?\,z_{max} : z$$

$$dist_{min} = \sqrt{(x - x_{bbox\_min})^2} + \sqrt{(y - y_{bbox\_min})^2} + \sqrt{(z - z_{bbox\_min})^2}$$

Clustering around the bounding box is quick and efficient.

Cluster Size Around Equipment



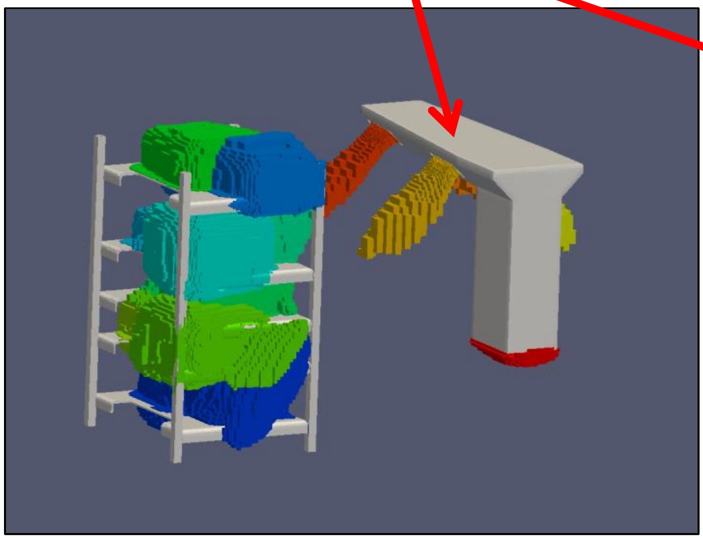


Larger weights allow the fixed location equipment-clusters to be smaller

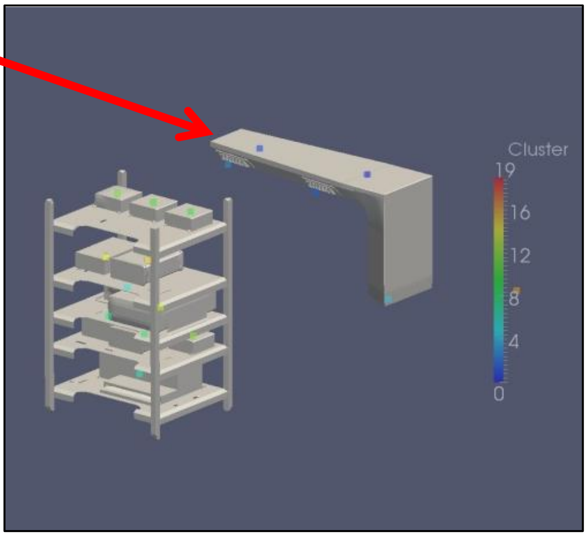
TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Fixing Clusters Around Inflow & Outflows

Important to Place Clusters at Inflow/ Outflows

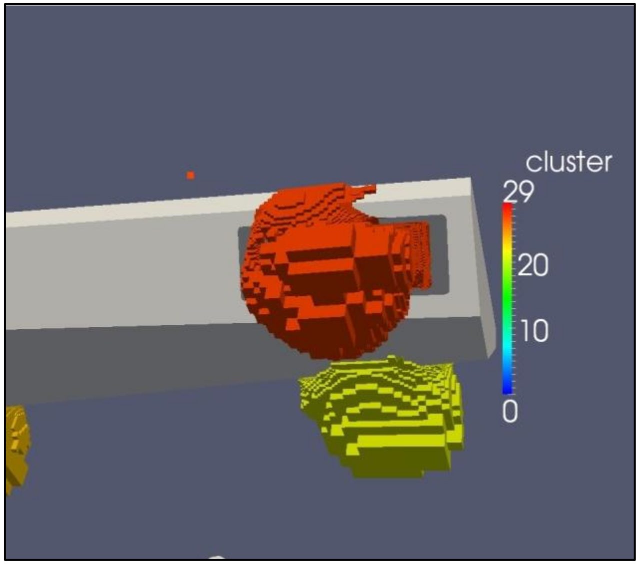


Cluster center locations



- Required to enforce flow network conservation of volume flow
- Usually an area of interest

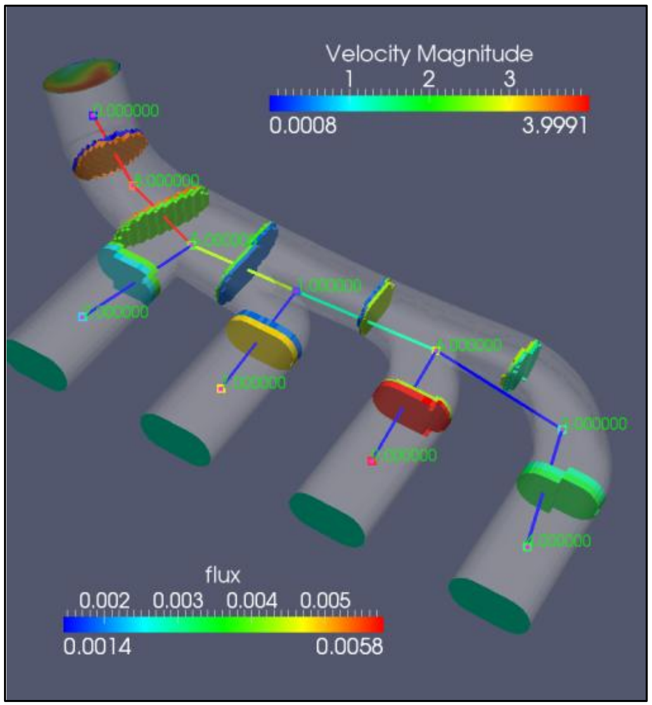
Bifurcation Problem at Inflow/Outflow



- Usually better to use centroid distance (as opposed to bounding box) from inflow/ outflow
- Added feature to specify cluster temperature, velocity at fixed nodes, to further avoid this

Validating Flux at The Interface Boundaries

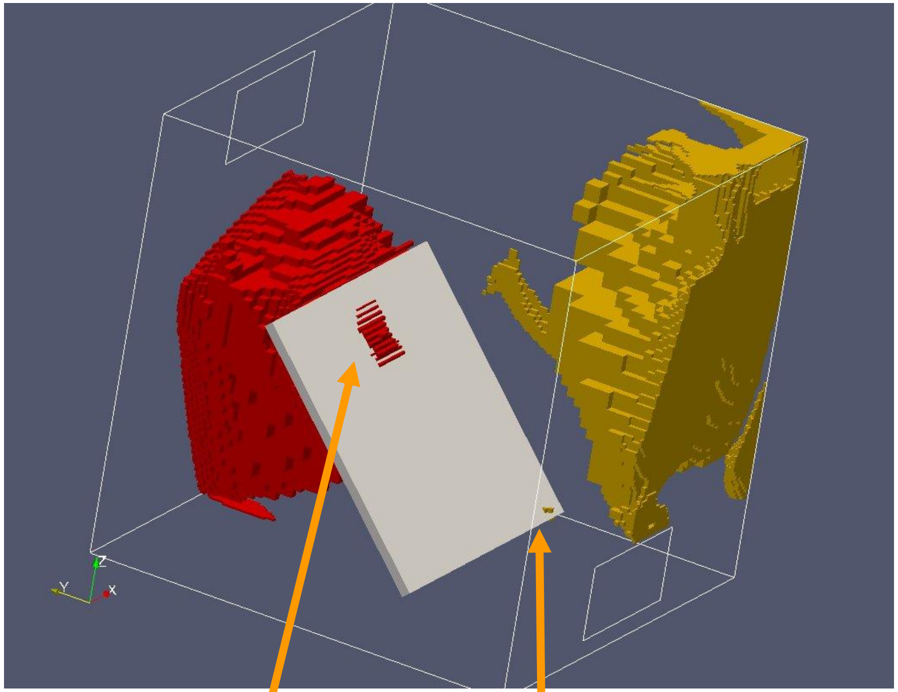
$$interface\ flux = \sum_{1}^{N_{faces}} a \cdot \frac{\vec{n}\cdot\vec{v}_1 + \vec{n}\cdot\vec{v}_1}{2}$$

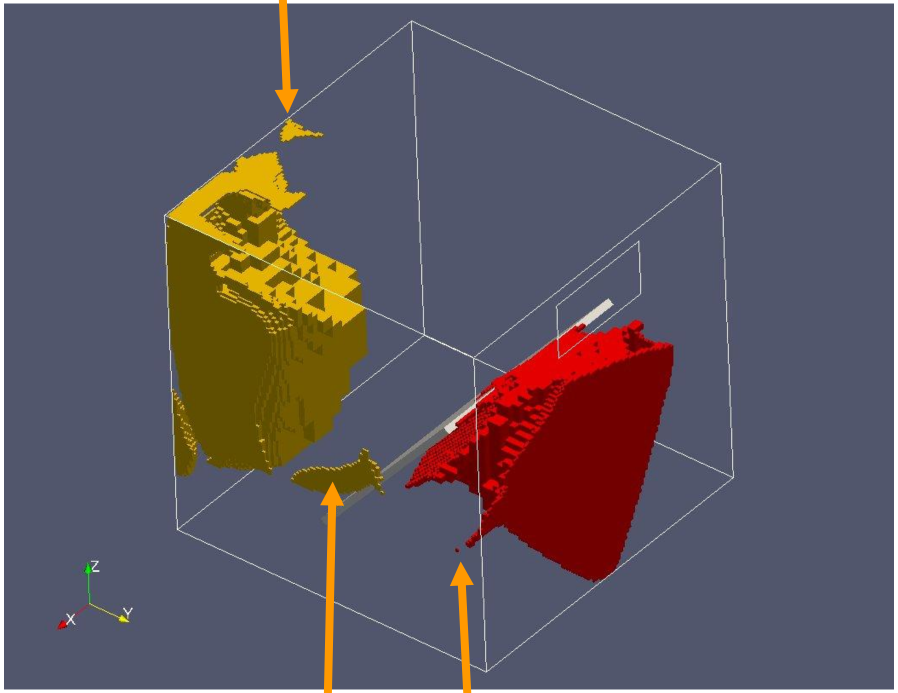


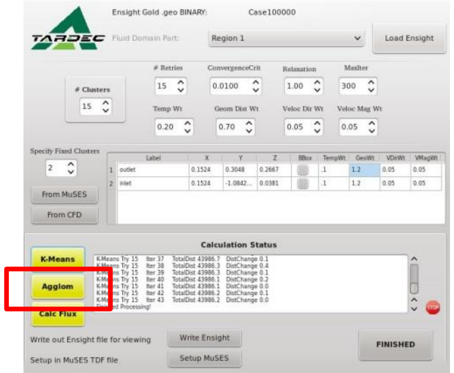
From Cluster	To Cluster	cfdMine m^3/s	StarCCM+ m^3/s	Error
8	0	0.005785	0.005798	0.23%
1	5	0.004345	0.004353	0.19%
6	1	0.002899	0.002908	0.33%
7	1	0.001445	0.001447	0.12%
2	5	0.001442	0.001445	0.19%
4	3	0.001447	0.001448	0.05%
3	6	0.001447	0.001448	0.05%
5	8	0.005788	0.005798	0.17%
9	6	0.001446	0.001445	-0.08%

This case is fairly idealized: Planes are perpendicular to flow, fine mesh, and no major temperature gradients.

Agglomeration – Dealing With Islands



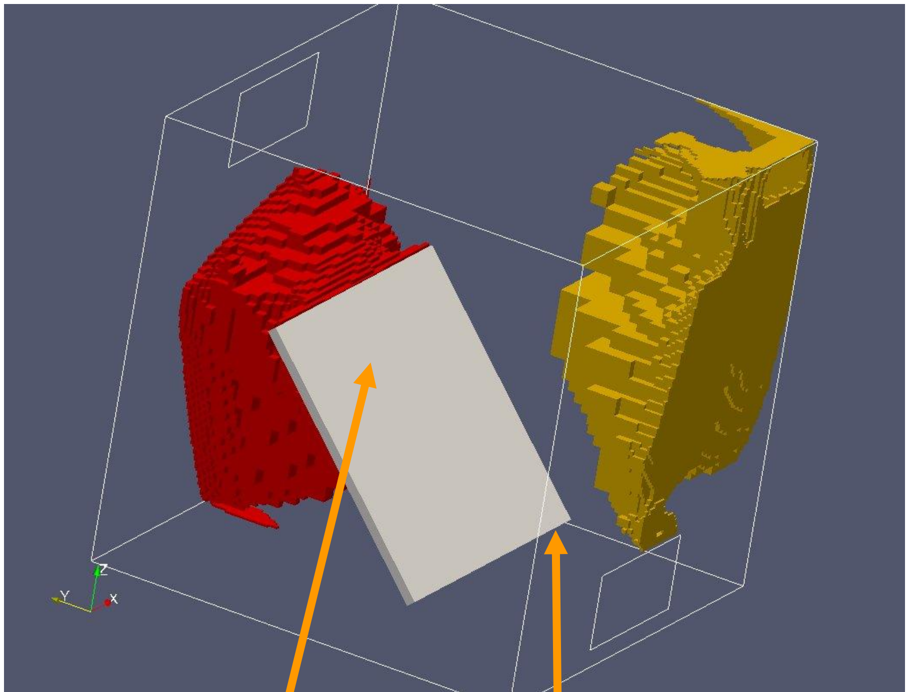


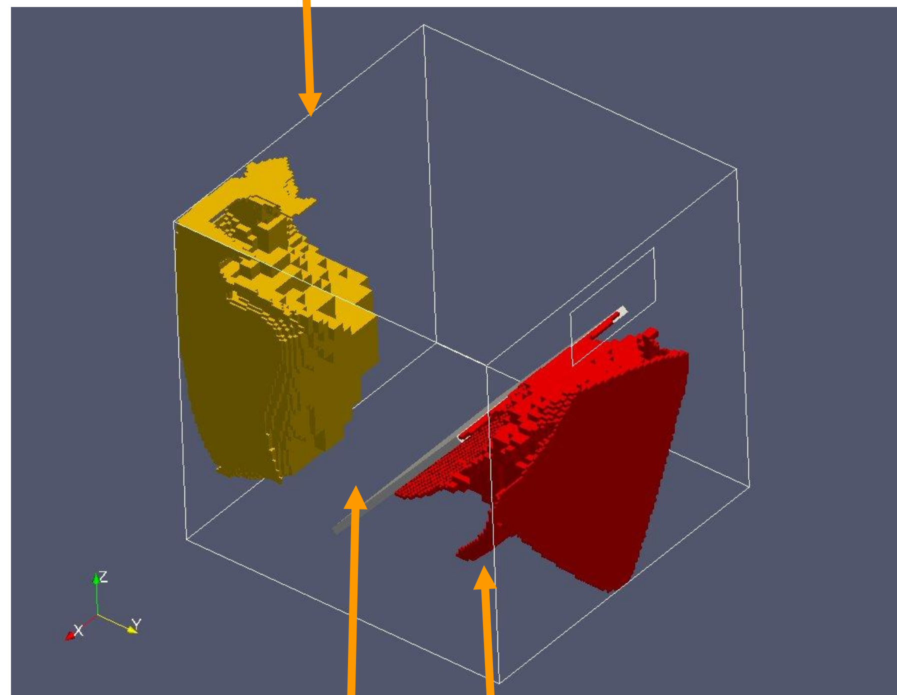


TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Agglomeration – Dealing With Islands

Island Eliminated





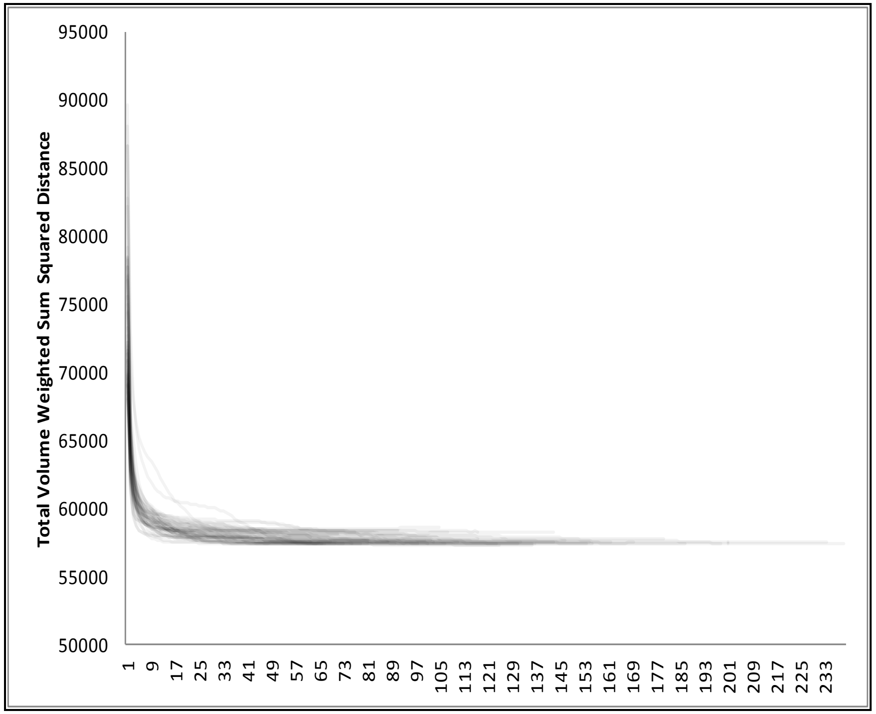
Island Eliminated

Island Eliminated

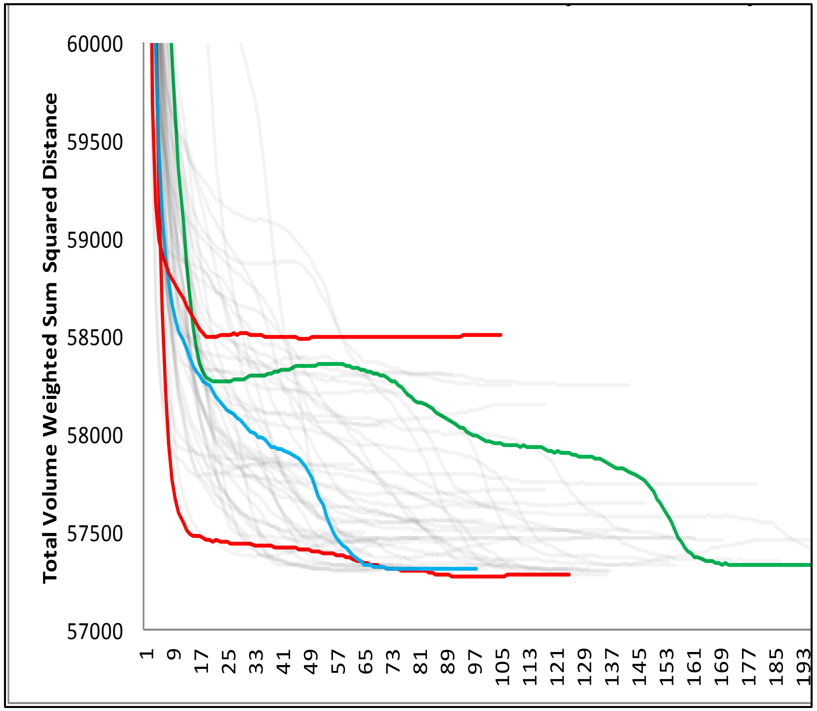
Island Eliminated

Island Eliminated

Convergence Plots



Typical k-means convergence plot for 50 random start trials

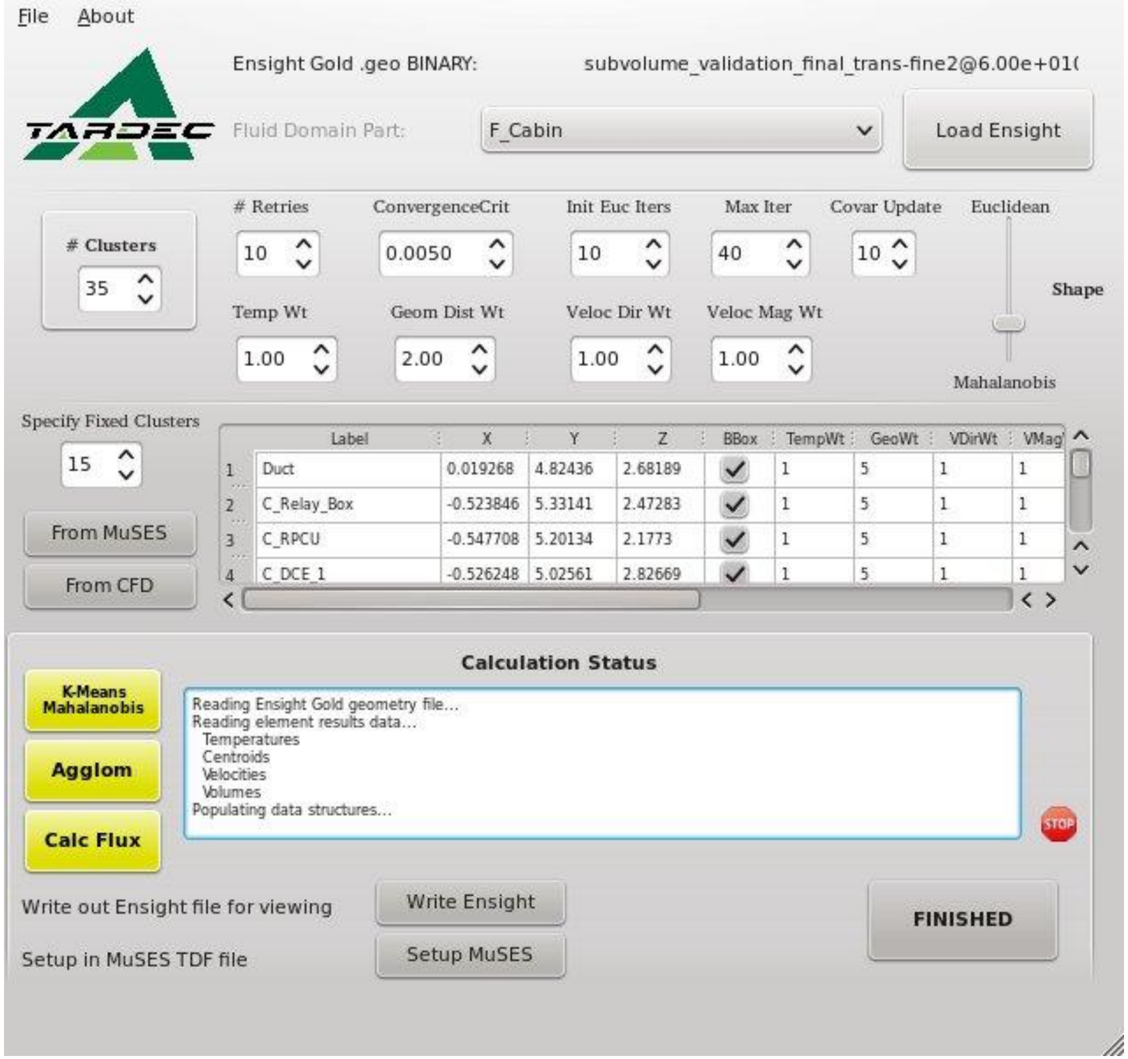


“zoomed-in” plot of 50 random-start k-means trials

Agenda

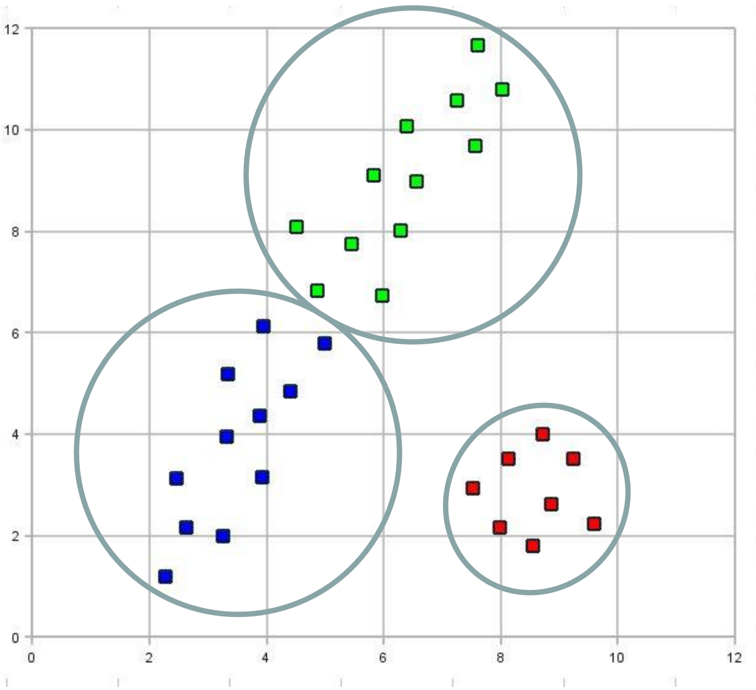
- About Rob E Smith
- Research Objectives
- Quick Introduction to Data Mining
- cfdMine Tool
- cfdMine Mahalanobis
- Validation Results
- Conclusions/ Next Steps

cfdMine Mahalanobis

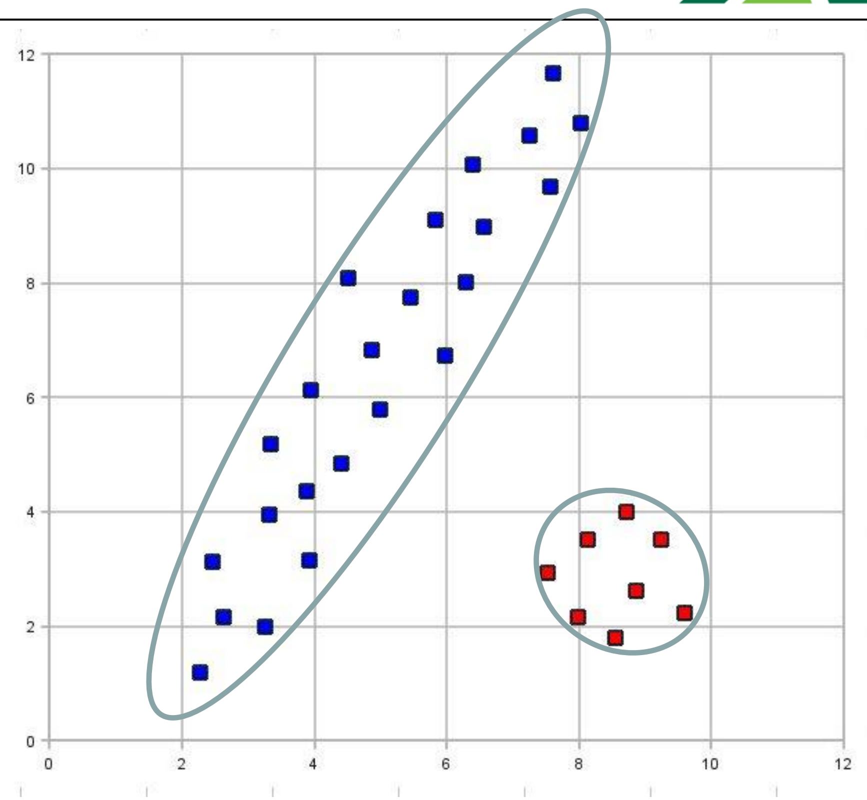


TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Mahalanobis Distance Metric



Euclidean Distance
(Standard K-Means)



Mahalanobis Distance

Euclidean distances fit clusters to hyperspheres.
Mahalanobis fits data to an oriented hyperellipsoid.

Mahalanobis Distance

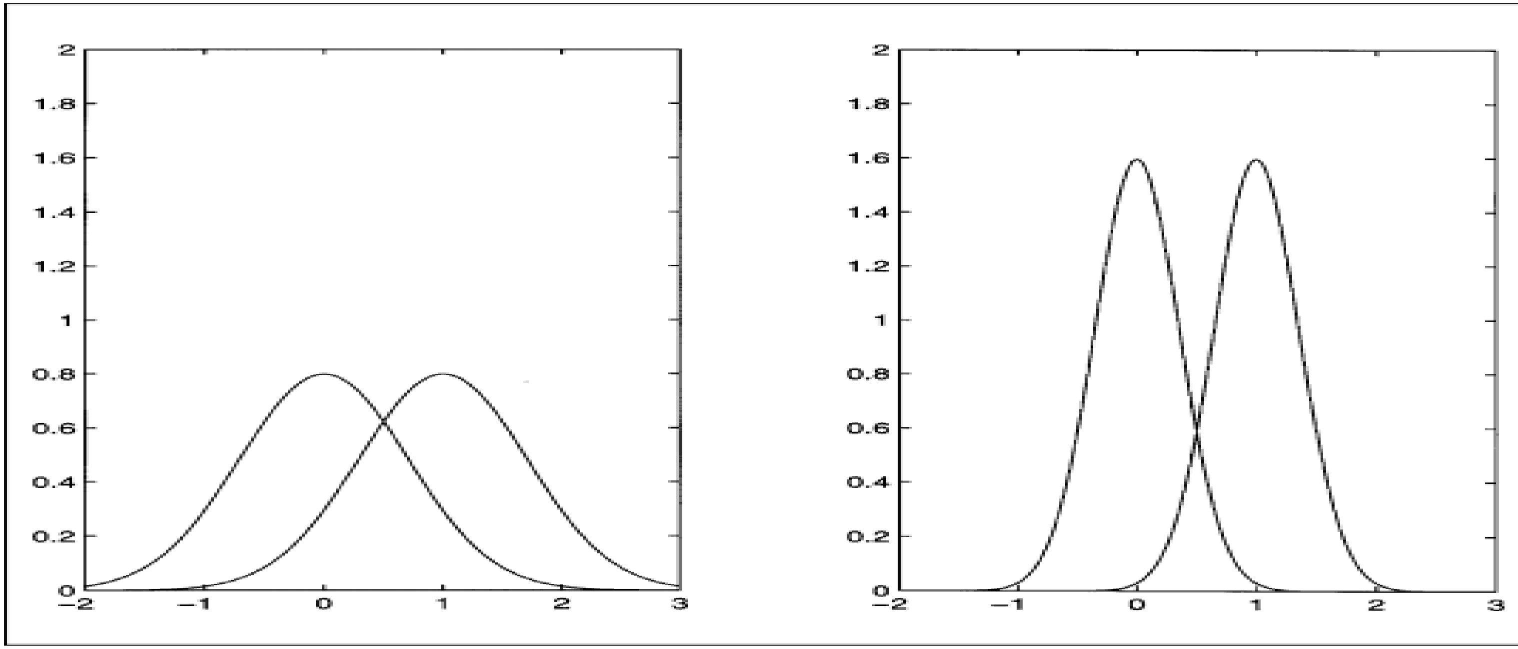
- $\Delta^2 = (x_i - \mu_k)^T \sum^{-1} (x_i - \mu_k)$, where

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{1L} \\ \sigma_{21} & \sigma_{22} & \sigma_{2L} \\ \sigma_{L1} & \sigma_{L2} & \sigma_{LL} \end{bmatrix}$$ covariance matrix

Note: An identity matrix results in Euclidean distance.

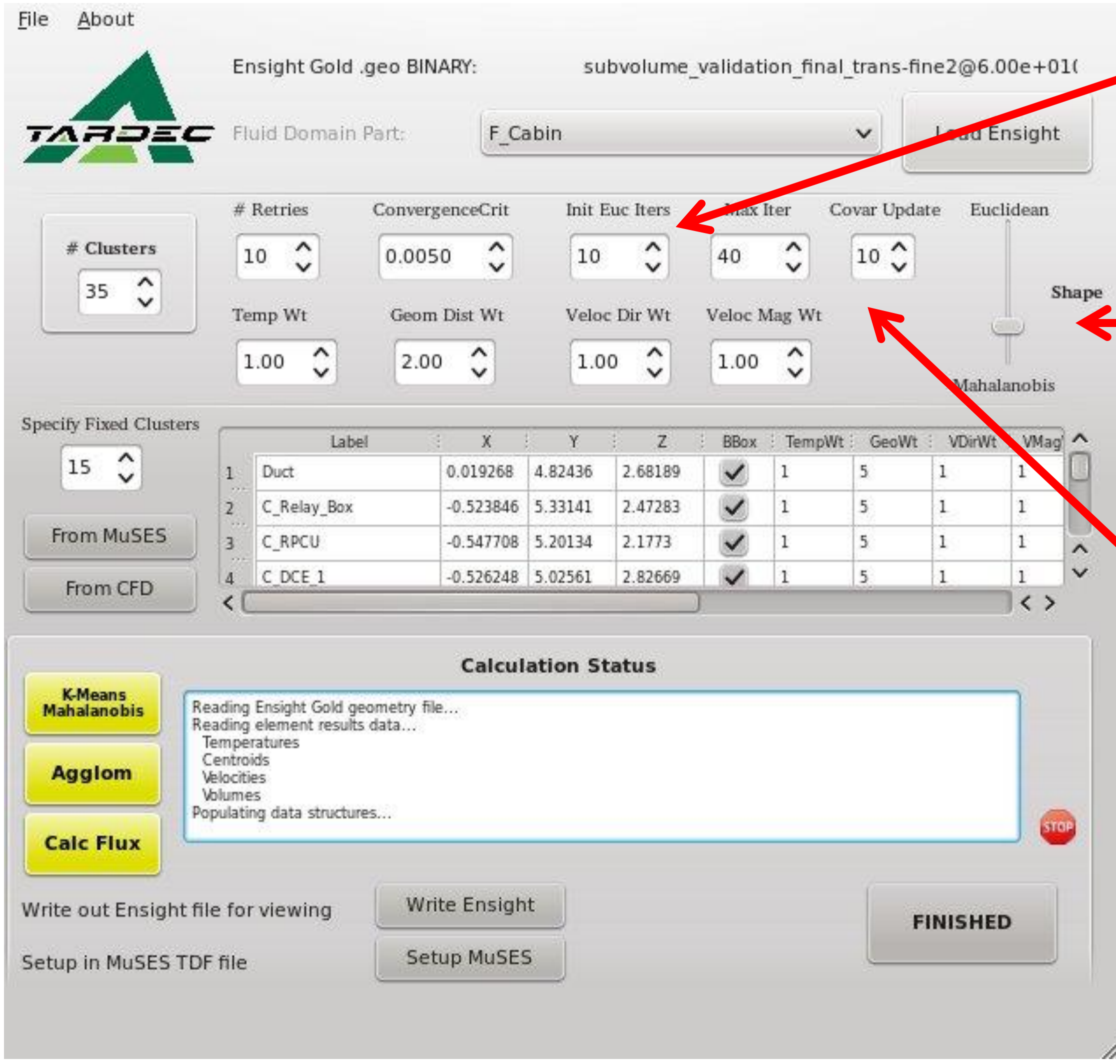
Mahalanobis cannot be used directly with K-means. Adds 36 unknowns and drops into non-optimal local minima even faster.

Example:Univariate Mahalanobis Distances



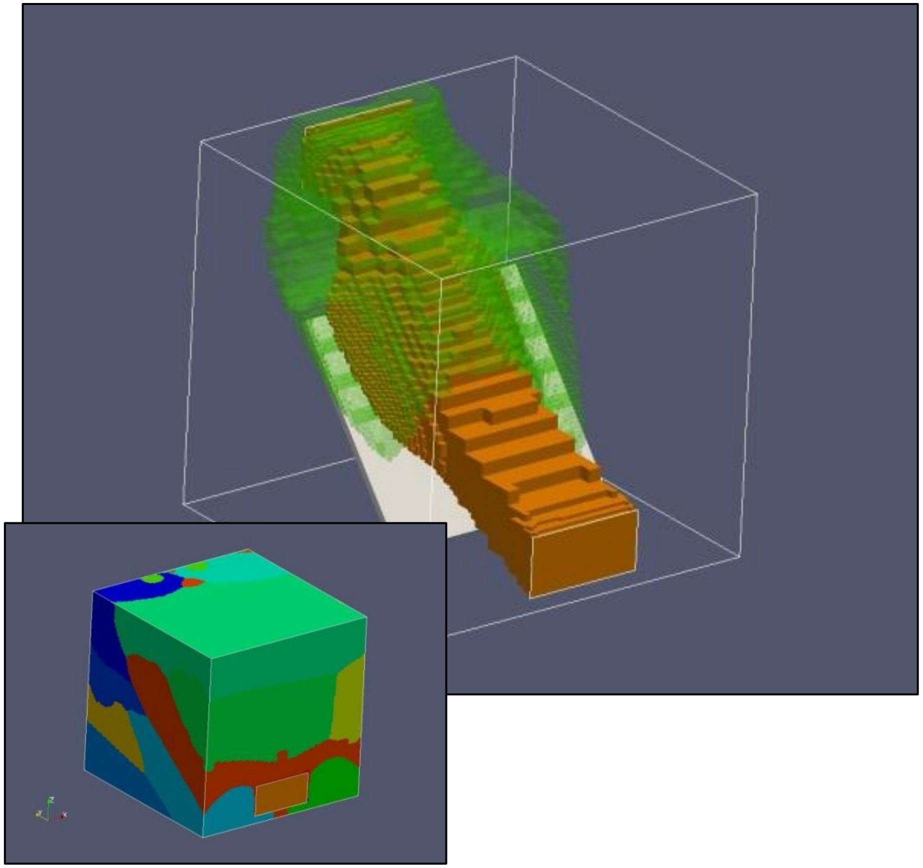
Euclidean distances are one in both cases, but Mahalanobis distance of Δ =1 and 2

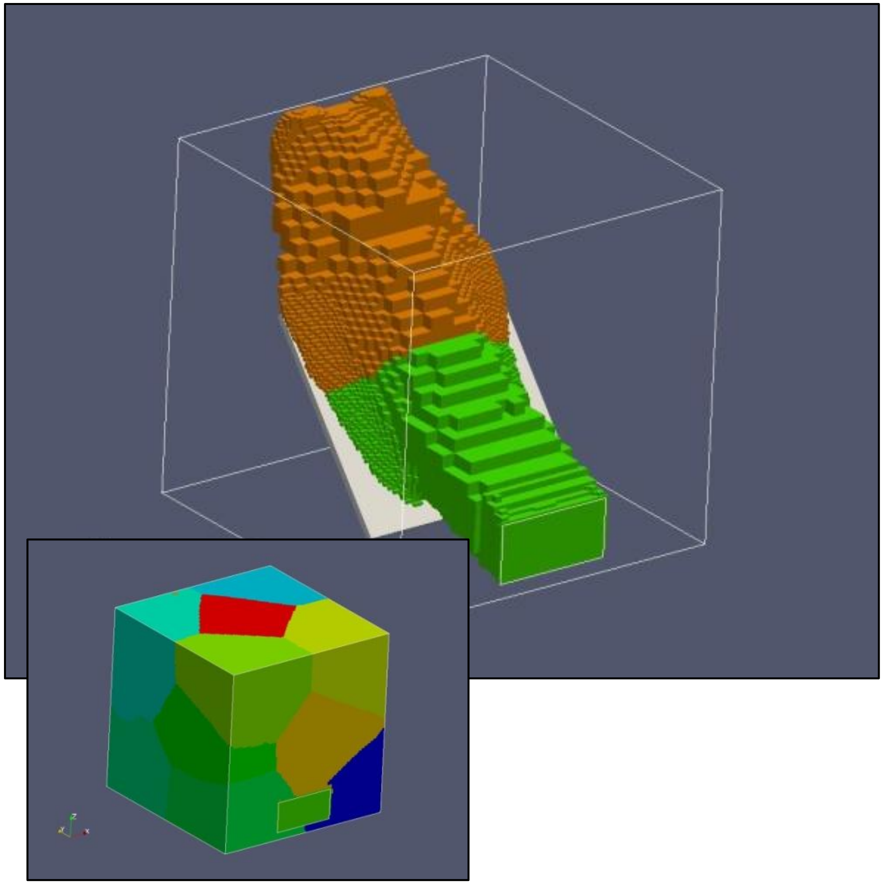
cfdMine Mahalanobis / New Algorithm



- Initial "m" pure Euclidean iterations
- Blend inv covariance w/ Euclidean(identity)
- Only update covariance "n" iterations

Results – Better Data Separation





Mahalanobis clusters

15 clusters, 75% Euclidean and 15% Mahalanobis

Pure Euclidean

15 clusters, 100% Euclidean

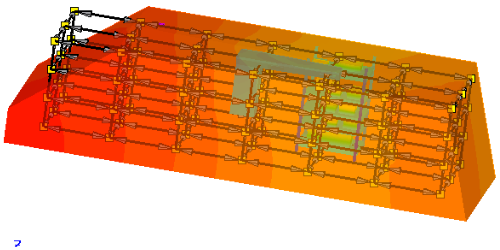
Mahalanobis resolves entire “jet” in Case 1, inlet to outlet.

Agenda

- About Rob E Smith
- Research Objectives
- Quick Introduction to Data Mining
- cfdMine Tool
- cfdMine Mahalanobis
- Validation Results
- Conclusions/ Next Steps

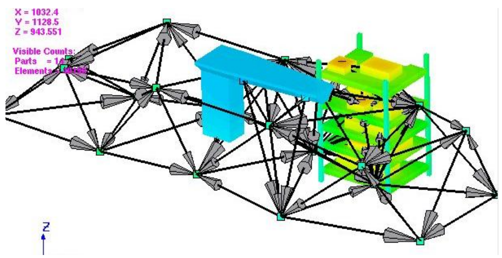
Validation Goal/ Method

- Subvolumed transient MuSES nodal network



  - Based on steady-state CFD
  - Transient 30 minute run

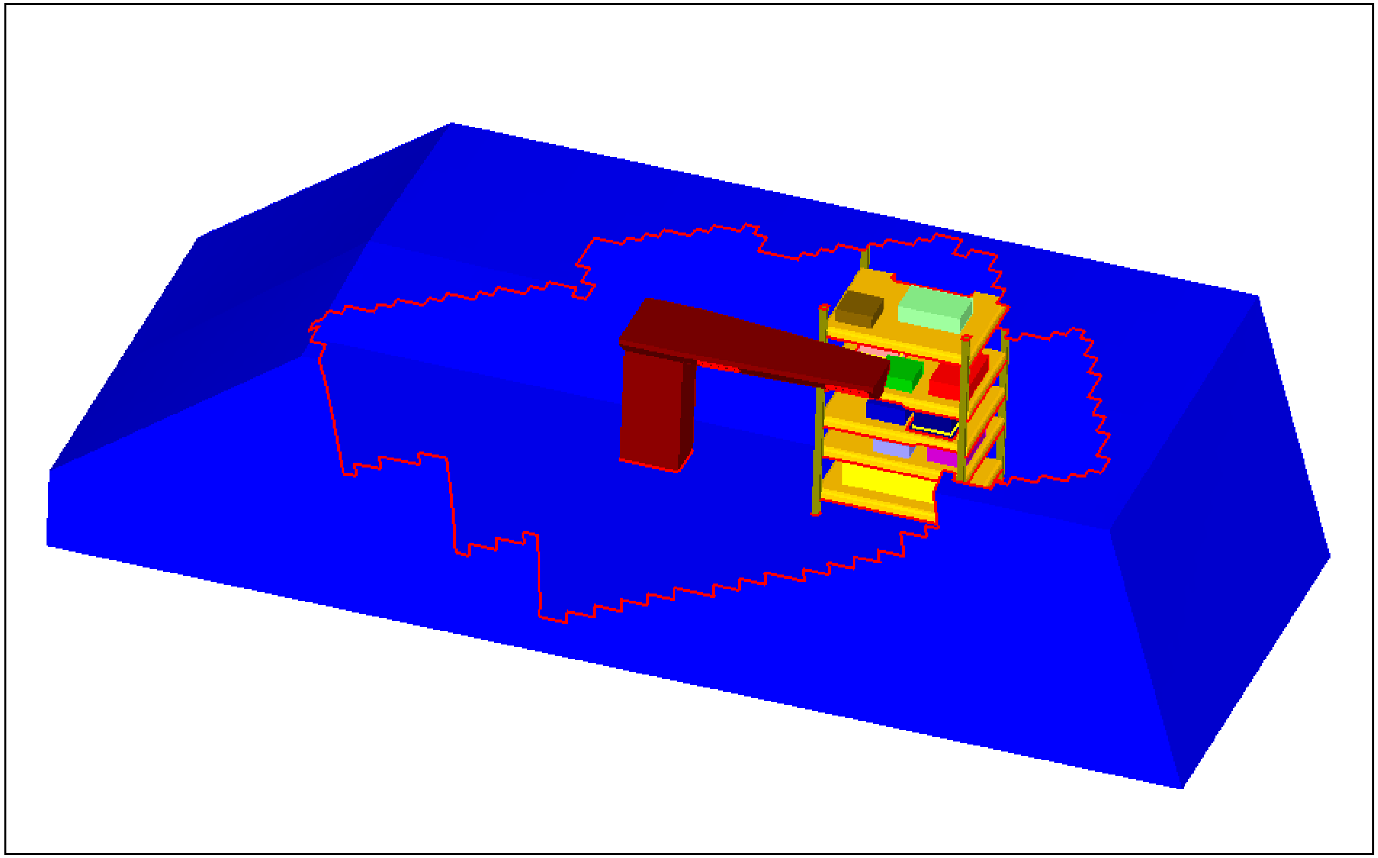
- Clustered transient MuSES nodal network



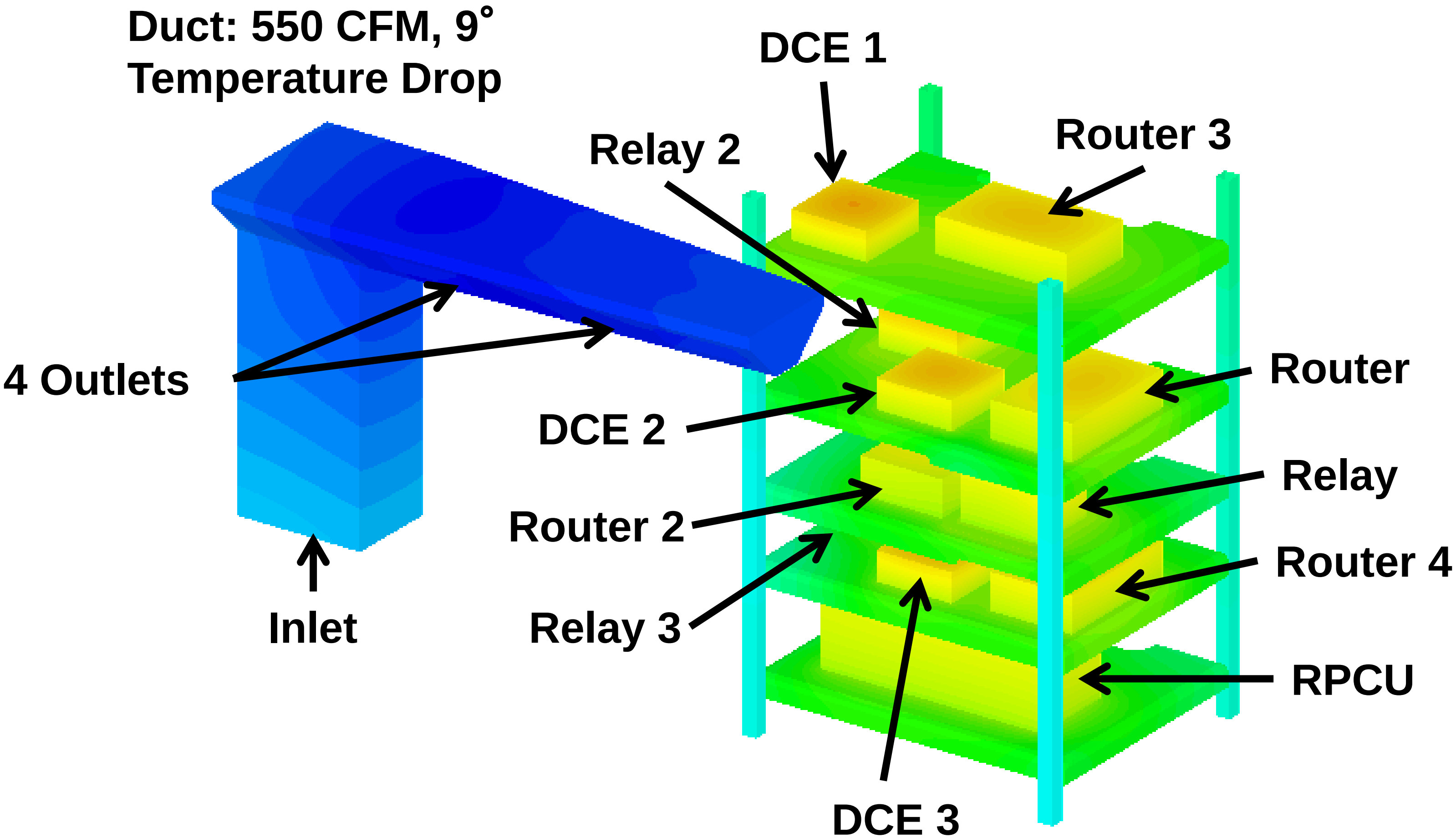
  - Based on steady-state CFD
  - Transient 30 minute run

- Compare to fully transient CFD for 30 minutes
  - Extract average cell temperatures from clusters each minute

Validation Study

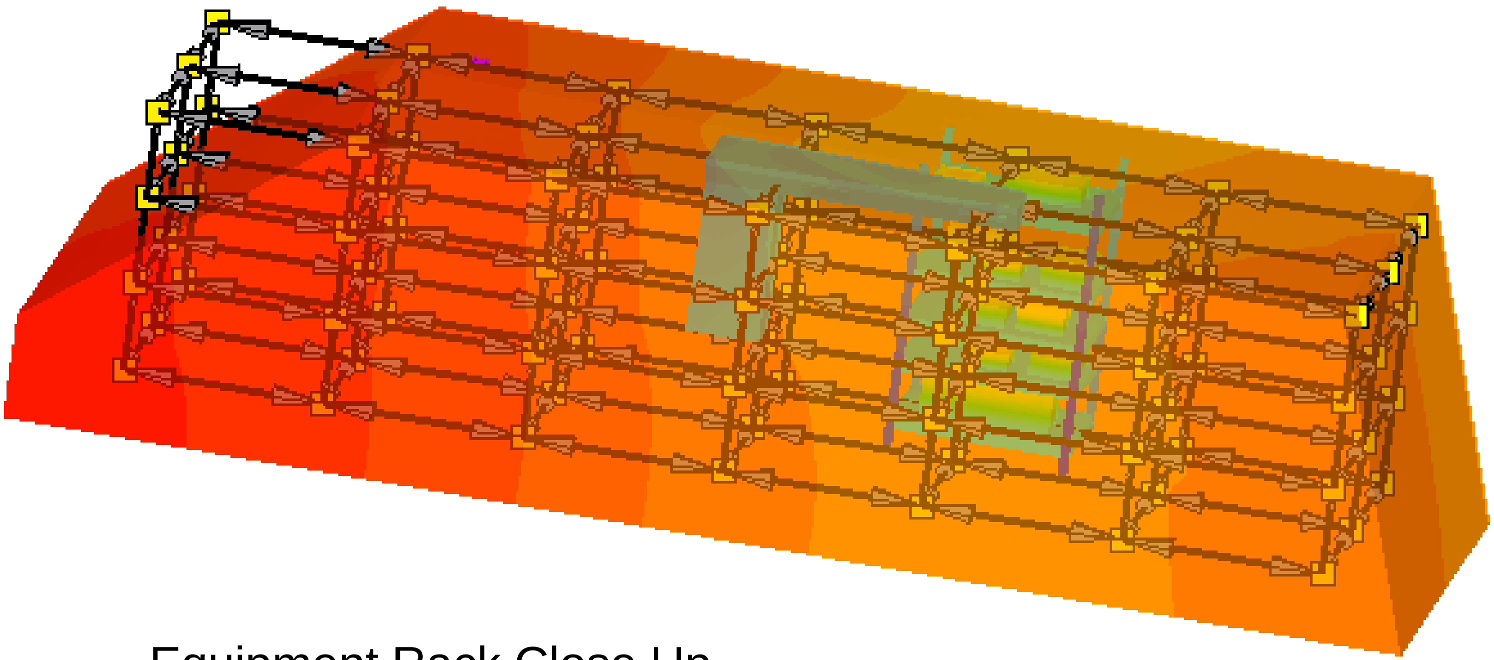


Equipment Rack Close-Up

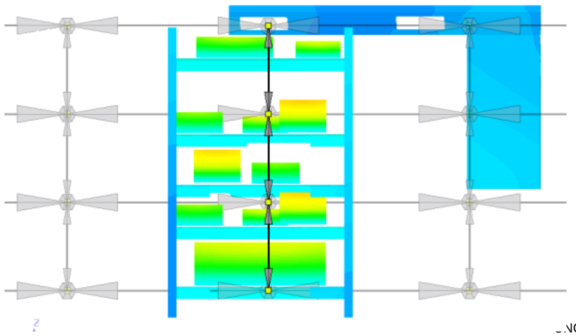


Subvolumed Nodal Arrangement

Subvolume nodal arrangement
72 Nodes Evenly Spaced



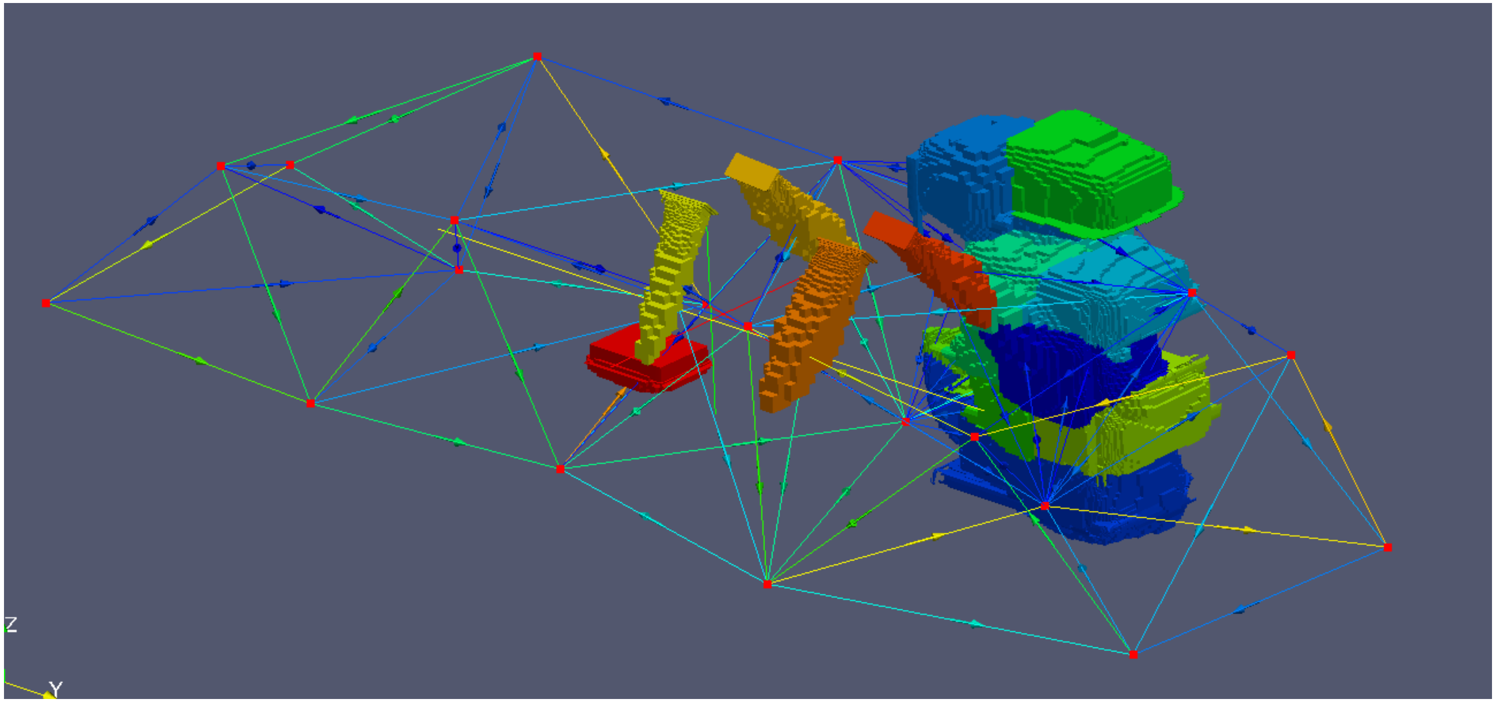
Equipment Rack Close Up



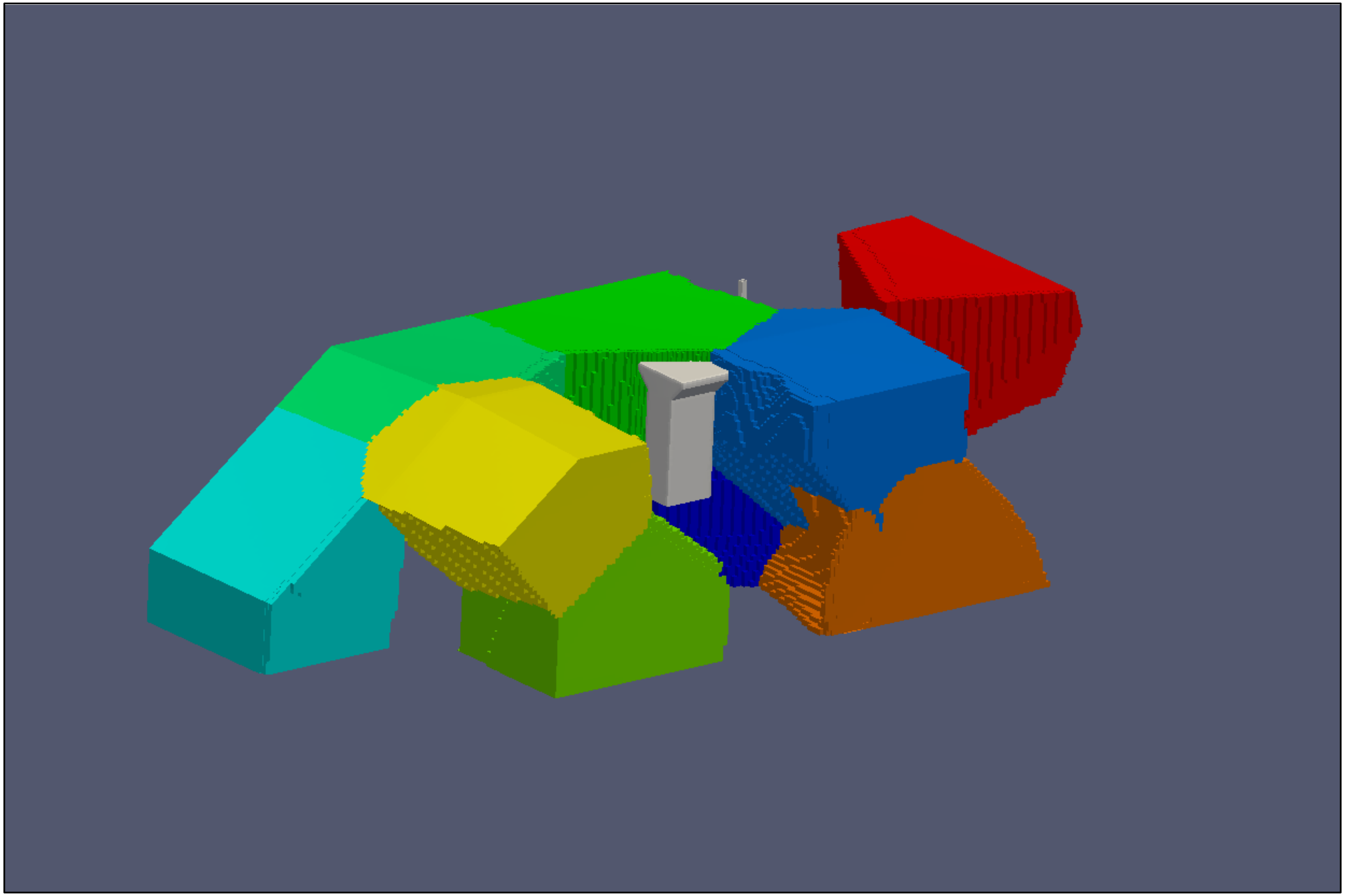
TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Clustering Nodal Arrangement

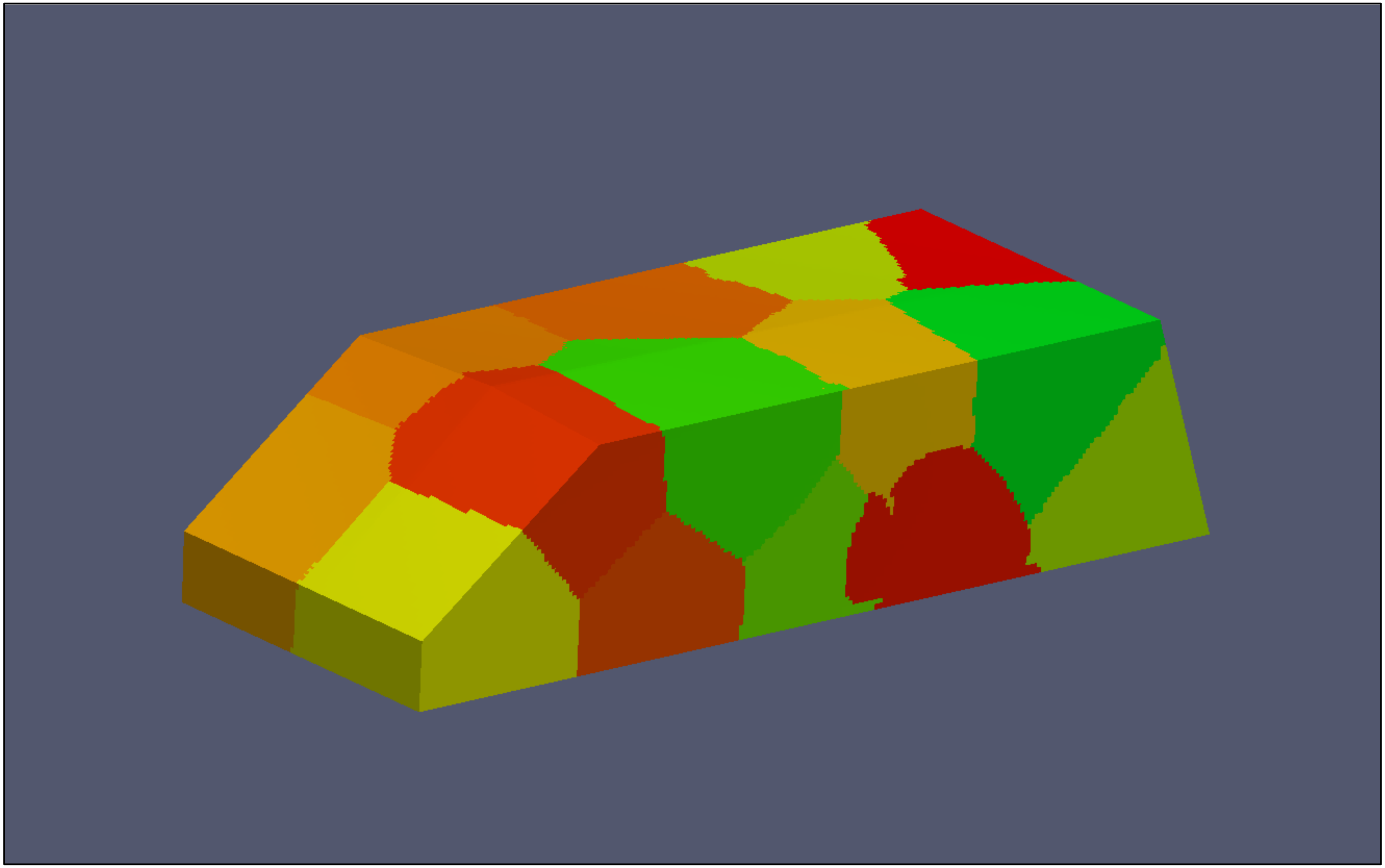
- 100% Euclidean
- Minimal time spent on clustering (not tailored)
- 35 Nodes
  - 10 around equipment
  - 5 around inlet/ outlets



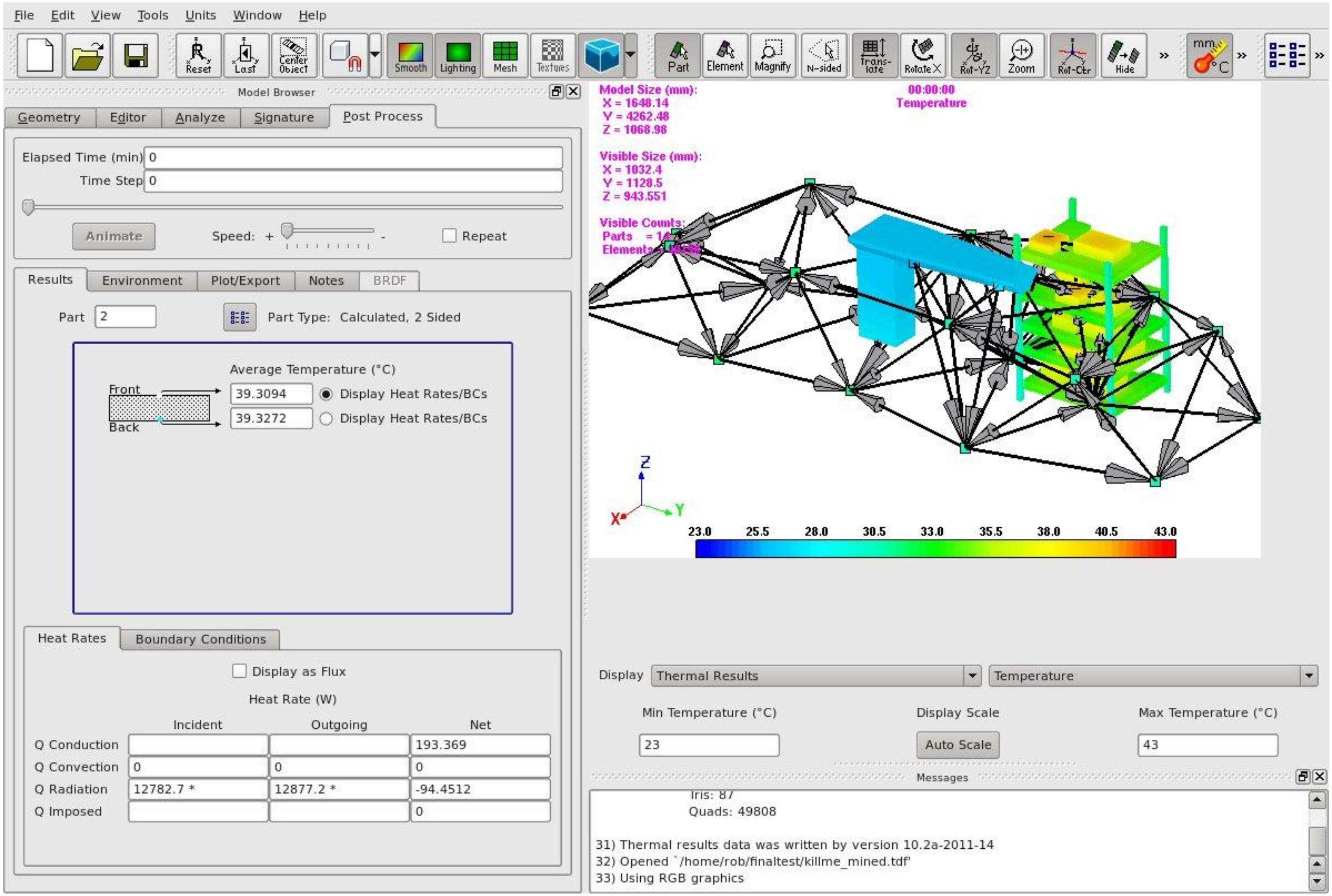
Validation Clustering View 2



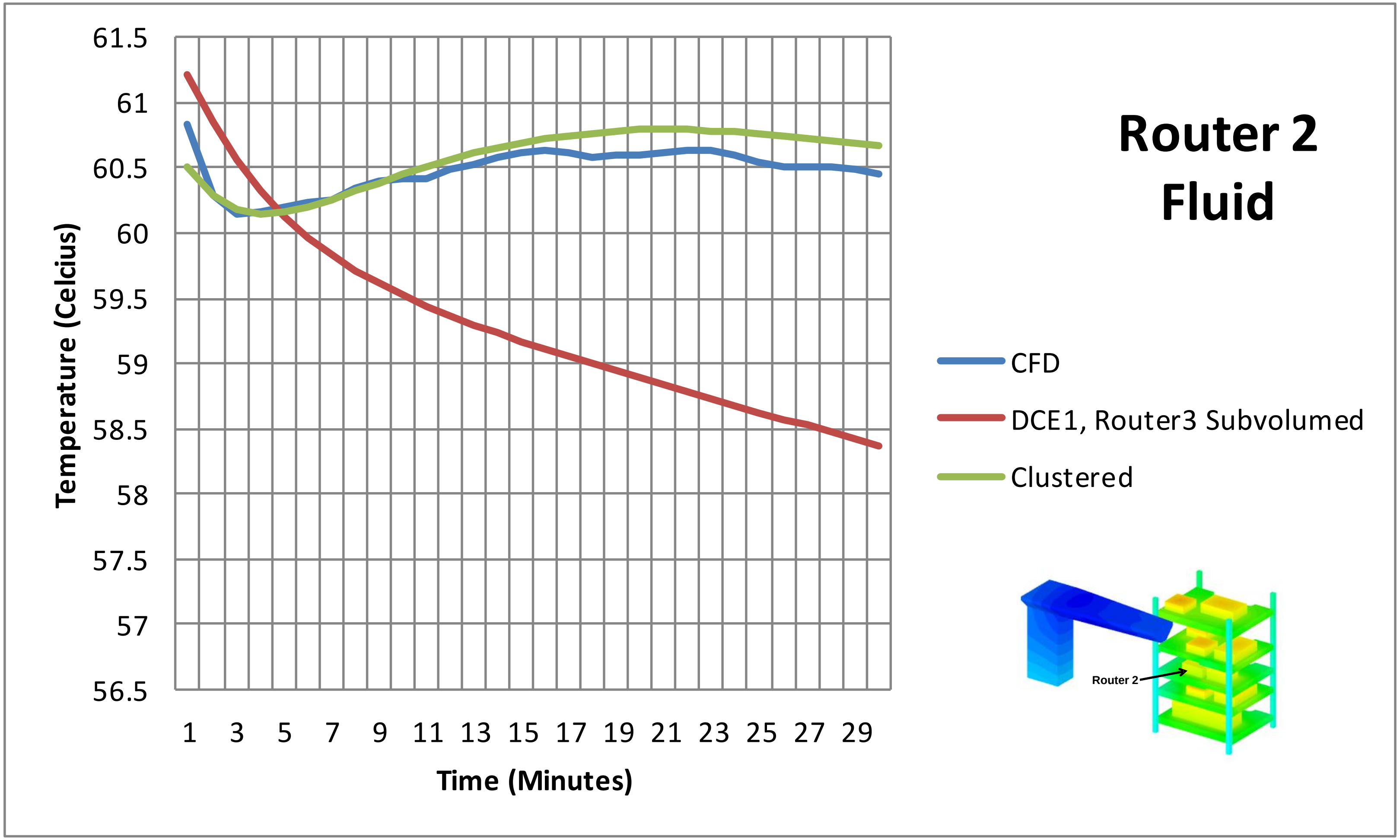
Validation Clustering View 3

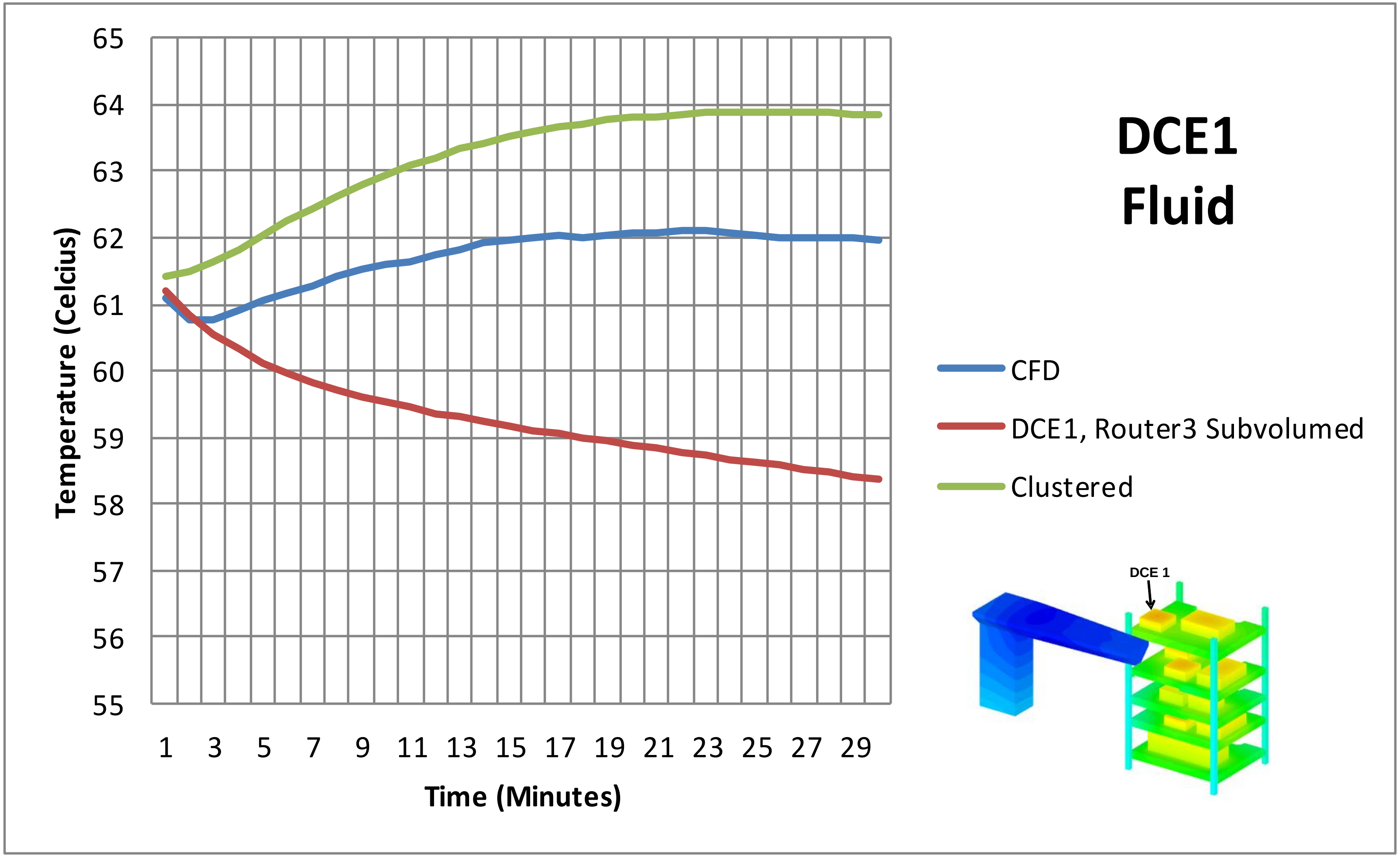


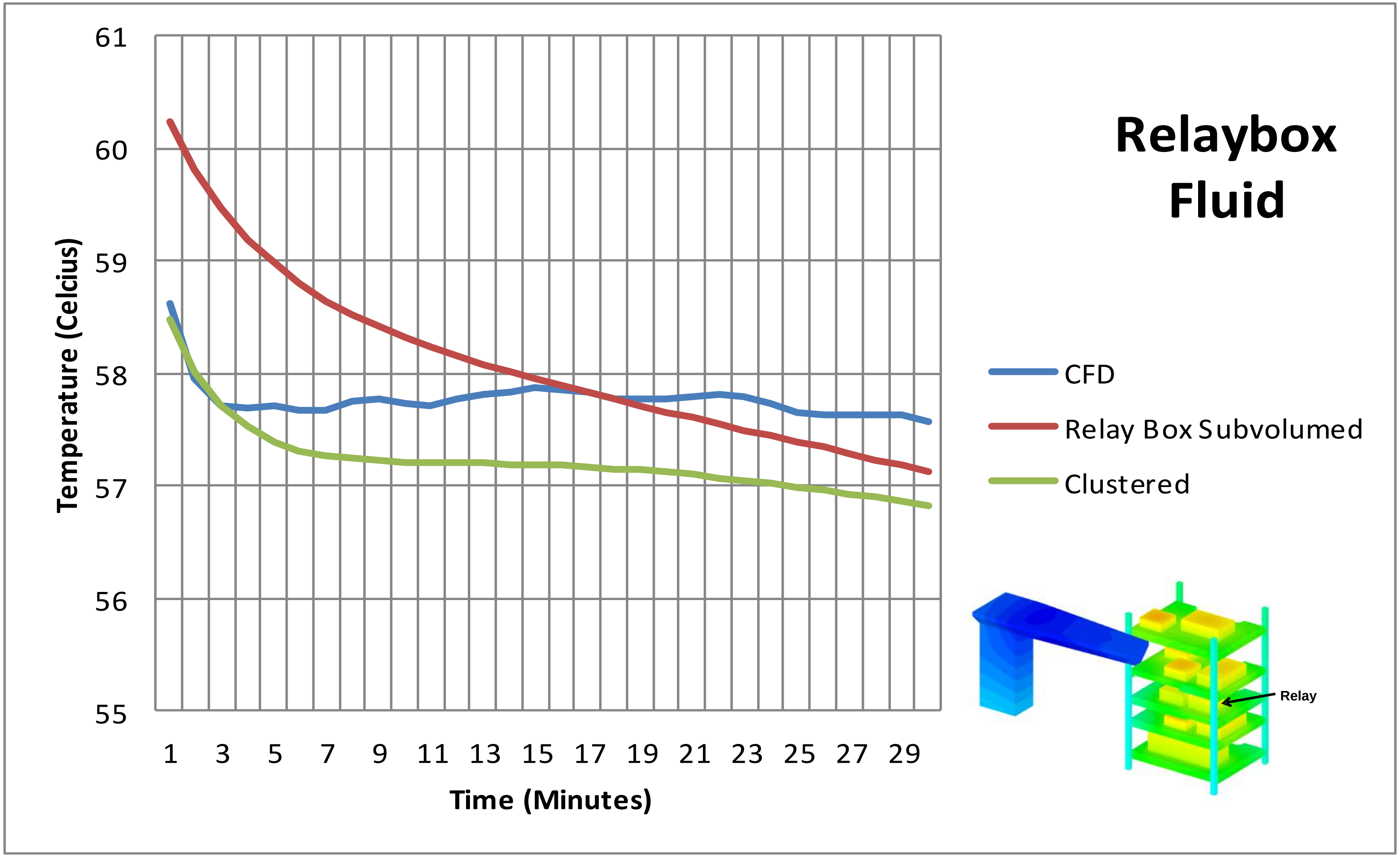
Clustered Network as Viewed in MuSES

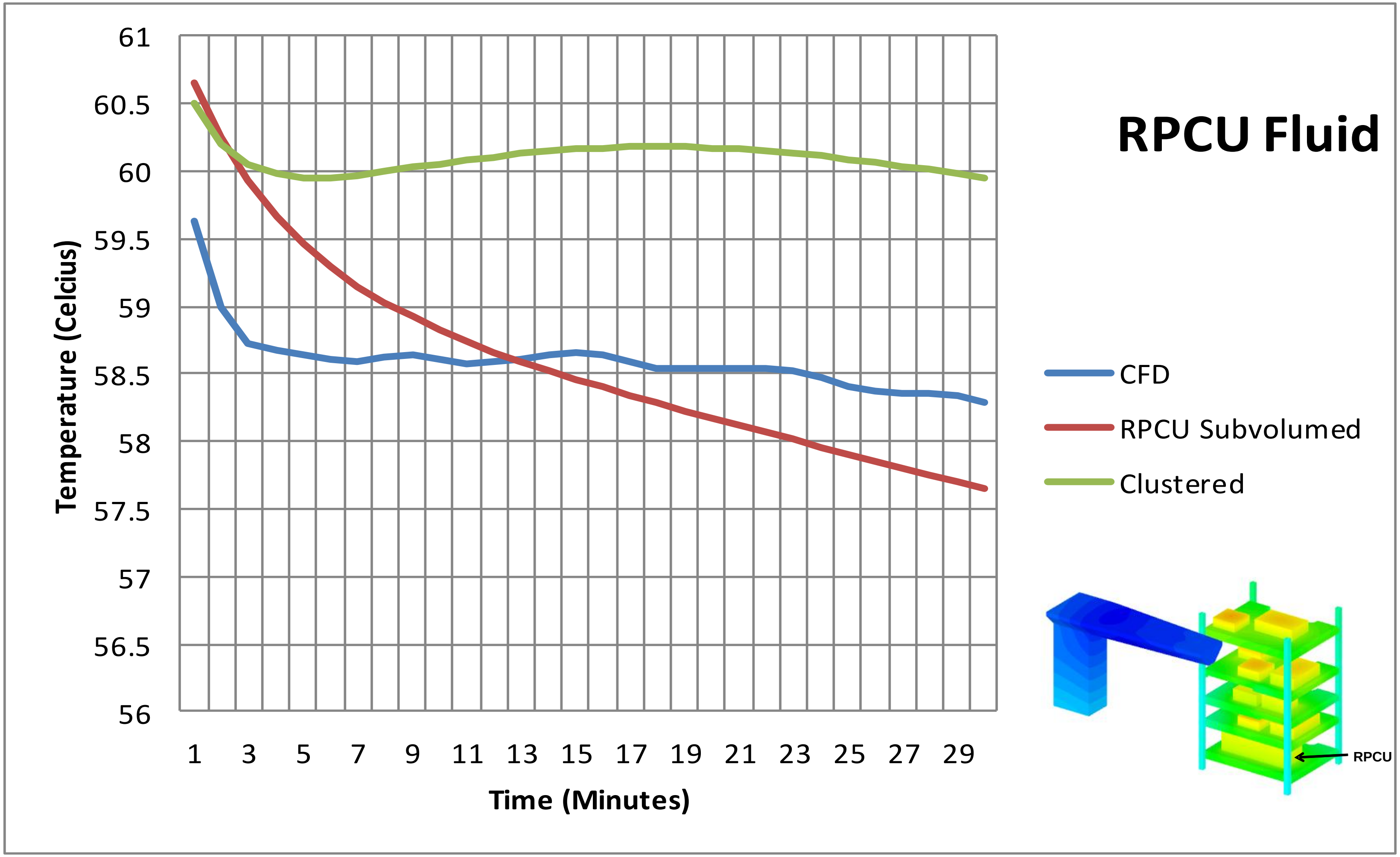


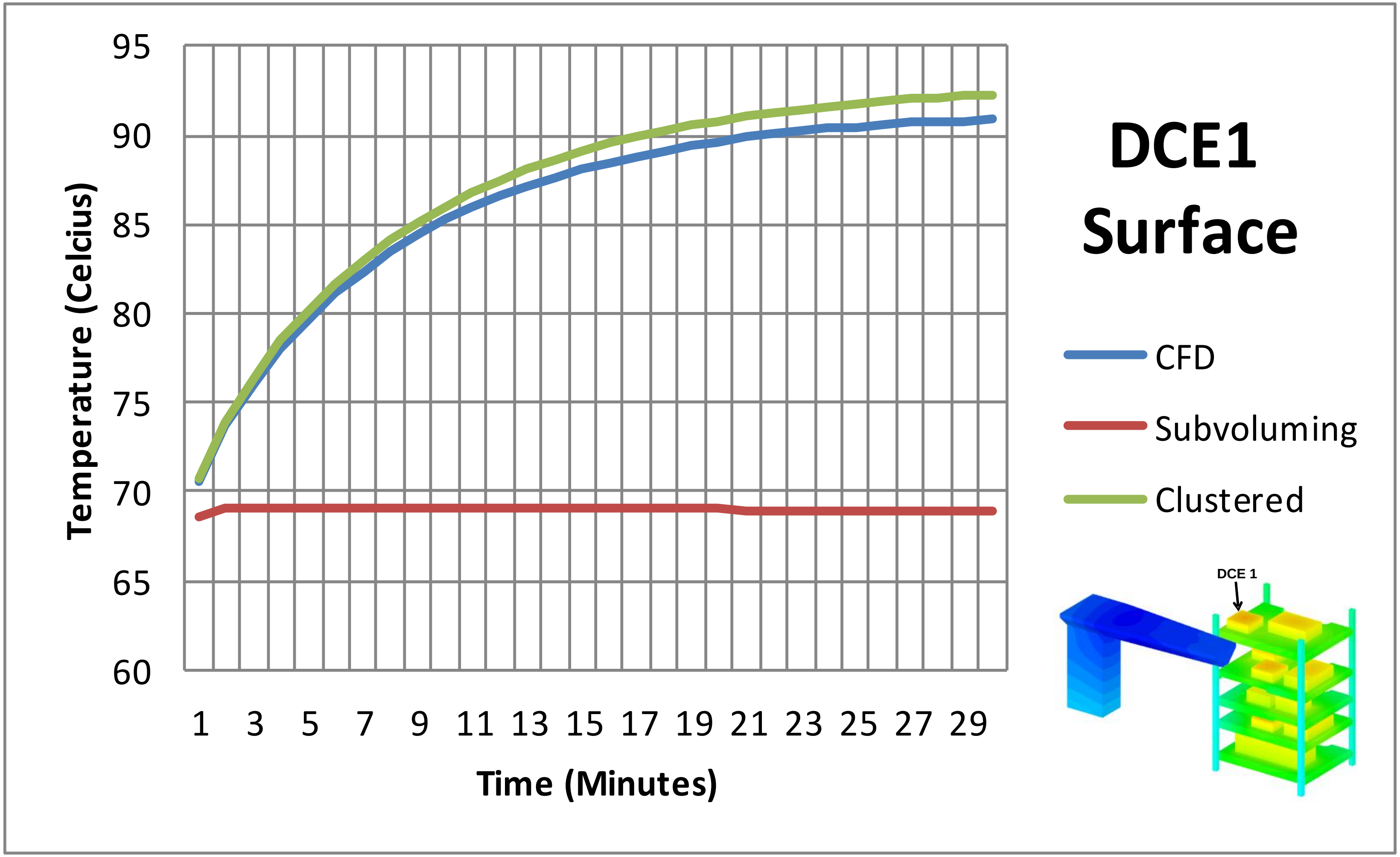
TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.


Router 2
Fluid
61.5
61
60.5
60
59.5
59
58.5
58
57.5
57
56.5
Temperature (Celcius)
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29
Time (Minutes)
CFD
DCE1, Router3 Subvolumed
Clustered
Router 2

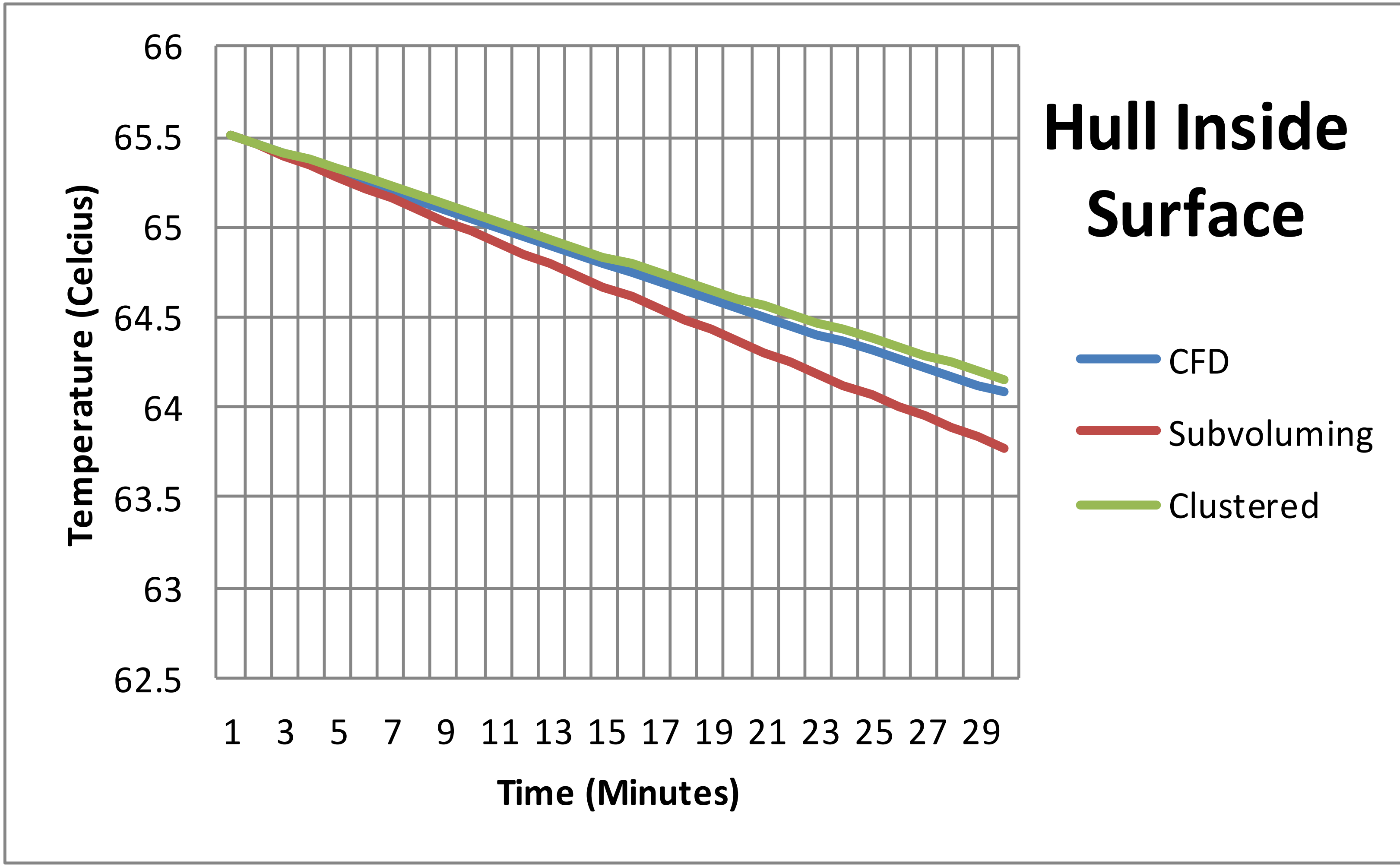

65
64
63
62
61
60
59
58
57
56
55
Temperature (Celcius)
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29
Time (Minutes)
DCE1
Fluid
CFD
DCE1, Router3 Subvolumed
Clustered
DCE 1


Relaybox
Fluid
61
60
59
58
57
56
55
Temperature (Celcius)
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
Time (Minutes)
CFD
Relay Box Subvolumed
Clustered
Relay

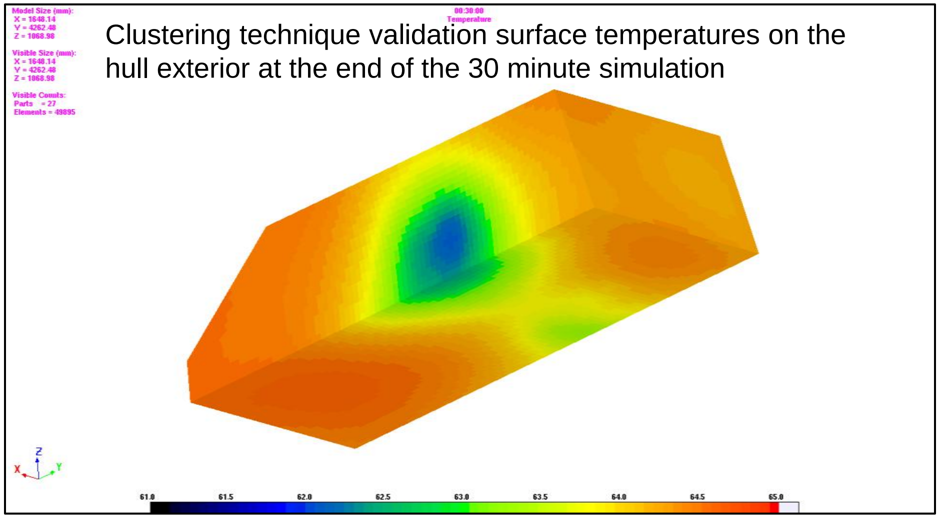

RPCU Fluid
Temperature (Celcius)
61
60.5
60
59.5
59
58.5
58
57.5
57
56.5
56
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
Time (Minutes)
CFD
RPCU Subvolumed
Clustered
RPCU



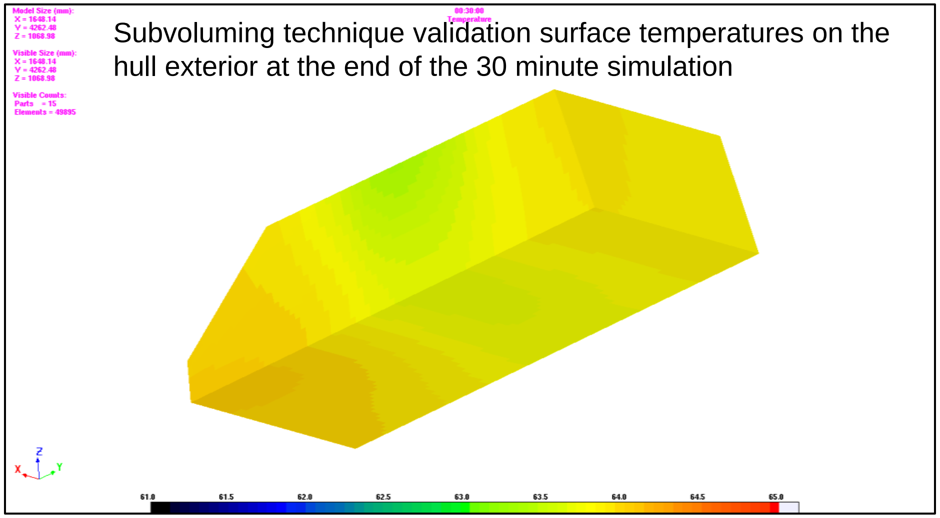
95
90
85
80
75
70
65
60
Temperature (Celcius)
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29
Time (Minutes)
DCE1
Surface
CFD
Subvoluming
Clustered
DCE 1


Hull Inside Surface
66
65.5
65
64.5
64
63.5
63
62.5
Temperature (Celcius)
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29
Time (Minutes)
CFD
Subvoluming
Clustered

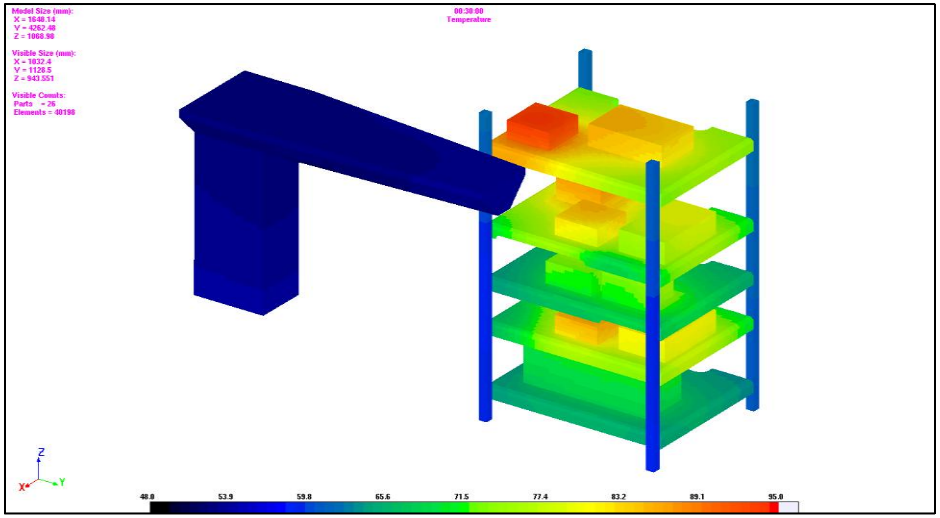
Demonstrated Need For Localized Convection



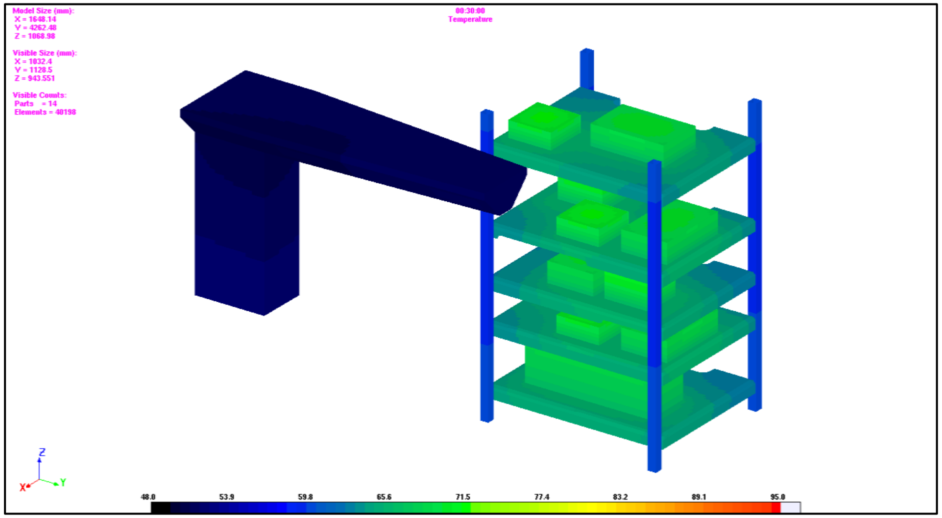
Clustering technique validation surface temperatures on the hull exterior at the end of the 30 minute simulation



Subvoluming technique validation surface temperatures on the hull exterior at the end of the 30 minute simulation



Clustering technique validation surface temperatures on the equipment rack at the end of the 30 minute simulation.



Subvoluming technique validation surface temperatures on the equipment rack at the end of the 30 minute simulation.

Remapping Problem

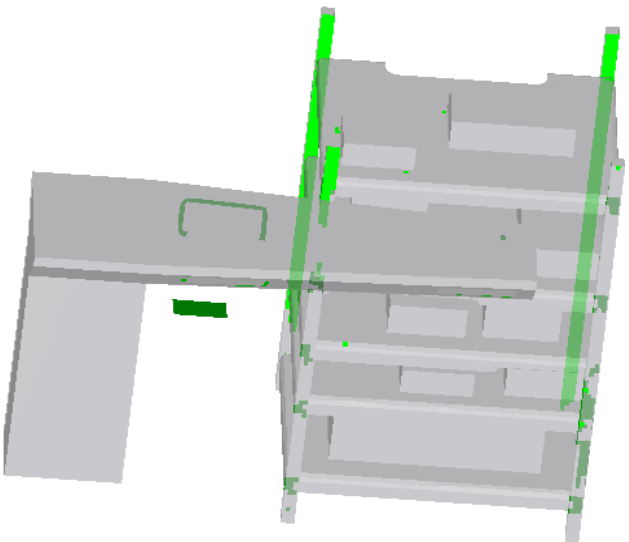
- Sometimes:

$$h_{local} = \frac{h_{film}(T_{film} - T_{wall})}{T_{local} - T_{wall}}$$

produces a negative h_{local}

Options:

- Set to zero
- Use absolute value <-Recommended



Validation Case Elements Which Have Negative Convection When Remapped

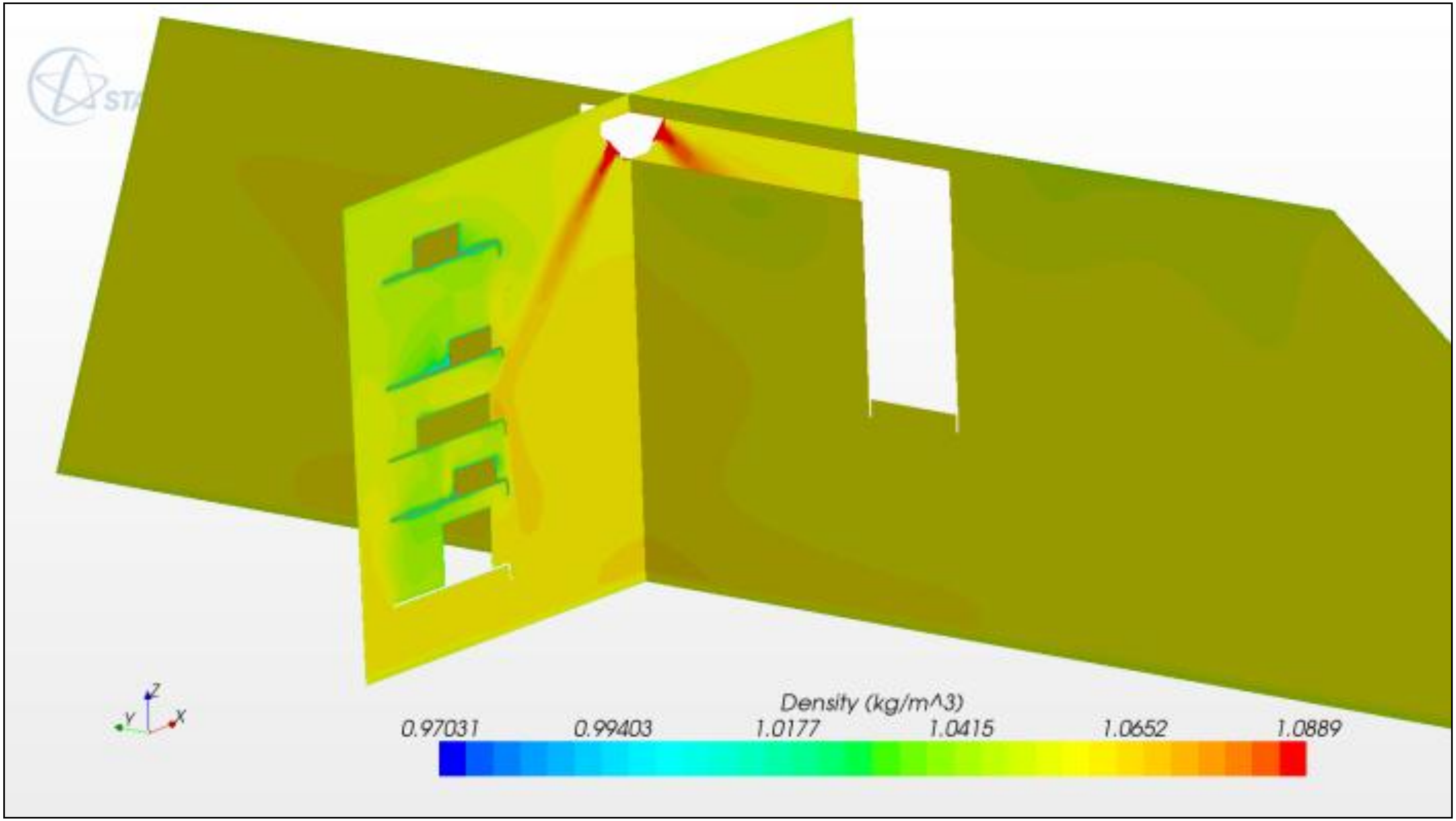
Conservation of Volume Problem

- It was noted the network is leaky for both subvoluming and cluster based networks.
  - Due to tolerances
  - Up to 10% imbalance seen on test case
  - Also should have been based on conservation of mass.

- Volume Balance Fix:

- Start with the first cluster and assess all incoming and outgoing advection links. If there is a source outside the domain, enforce this in the calculation as well.
- Divide the imbalance in half and apply half to the incoming and half to the outgoing flow to make them balance to zero (or reduce by a relaxation factor). Do this proportional to each advection link divided by the total incoming or outgoing flow.
- Evaluate each cluster according to step 2 successively until each cluster exhibits no further imbalance.

Density Plot



Agenda

- About Rob E Smith
- Research Objectives
- Quick Introduction to Data Mining
- cfdMine Tool
- cfdMine Mahalanobis
- Validation Results
- Conclusions/ Next Steps

Conclusions

- The new overall process of clustering a steady-state CFD domain into a transient lumped fluid network simulation performs well
- It is absolutely critical to conserve mass within the fluid network
- Remapping CFD fluid film convection to localized fluid nodes is critical to an accurate solution
- Mahalanobis distance is better than Euclidean, but too computationally intensive for the present constraints
- Clustering is an effective method to track temperatures around specific equipment or locations
- Clustering provides a beneficial new way to visualize flowfields
- Subvoluming should be used with extreme caution

New Contributions to the Field

- First to implement clustering to simplify a CFD network into a simplified network
- First known volume weighted clustering of a CFD domain.
- First known remapping of convection coefficients were remapped to localized nodes in a network
- First use of fixed nodes where the nodes were fixed at areas of interest during clustering for the purpose of a simplified nodal network.
- Proposed new methods to stabilize the Mahalanobis distance and make it usable.
- Showed several new ways to visualize flow data

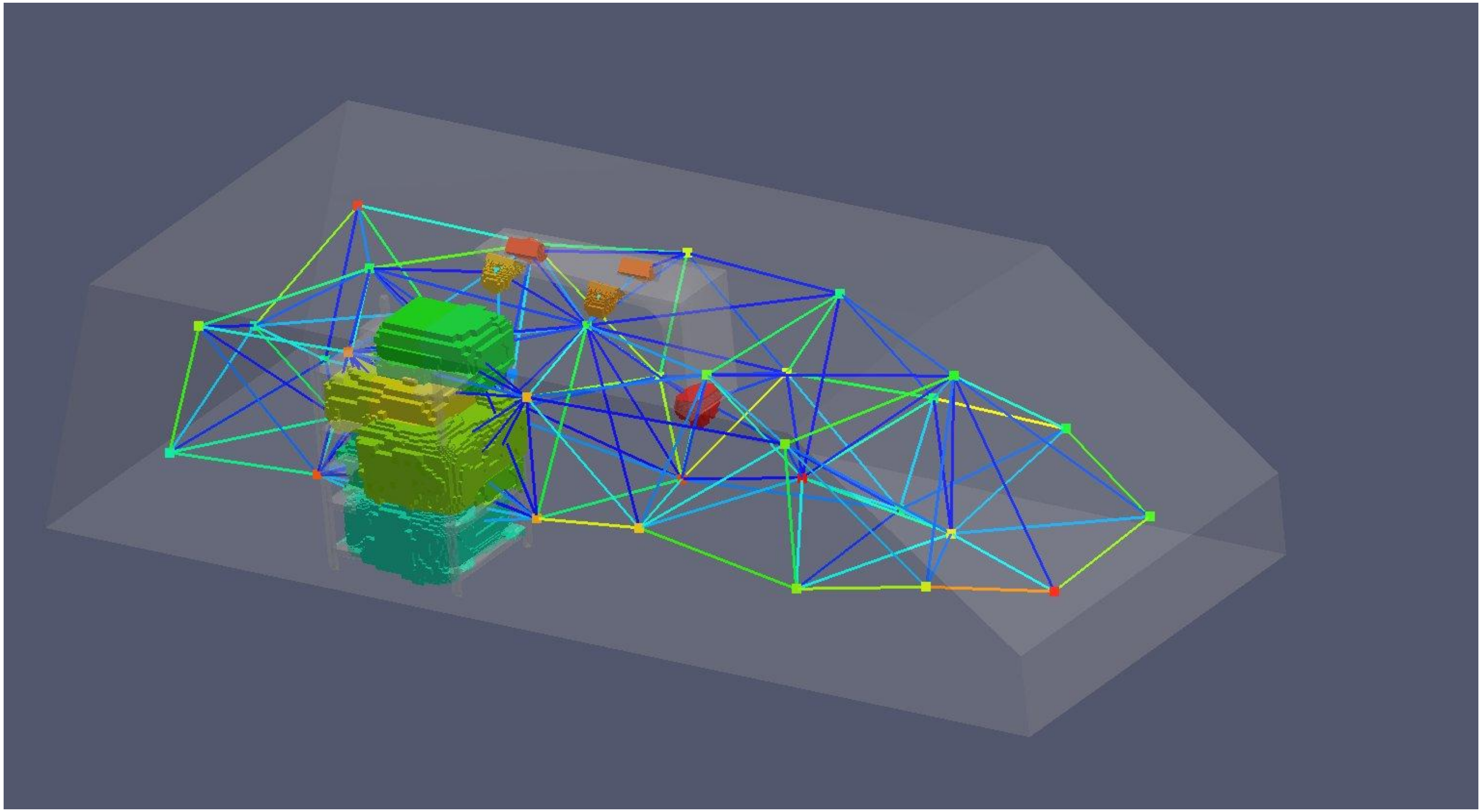
Future Work

- Implement conservation of mass
- Investigate use of unidirectional flux linkages
- Look into enhanced subvoluming
  - Control volumes around equipment
  - Localized convection
- Investigate importance of transient versus steady-state convection
- Convection "view factors"
  - Solve a steady-state CFD solution
  - Release particles from all wall elements and boundary elements (perform a particle trace)
  - Account for the mixing and the fraction of volume of influence each wall-released particle
  - Ultimately track the influence each wall element / boundary has on each other element/ boundary


LOONEY TUNES
QUESTIONS?

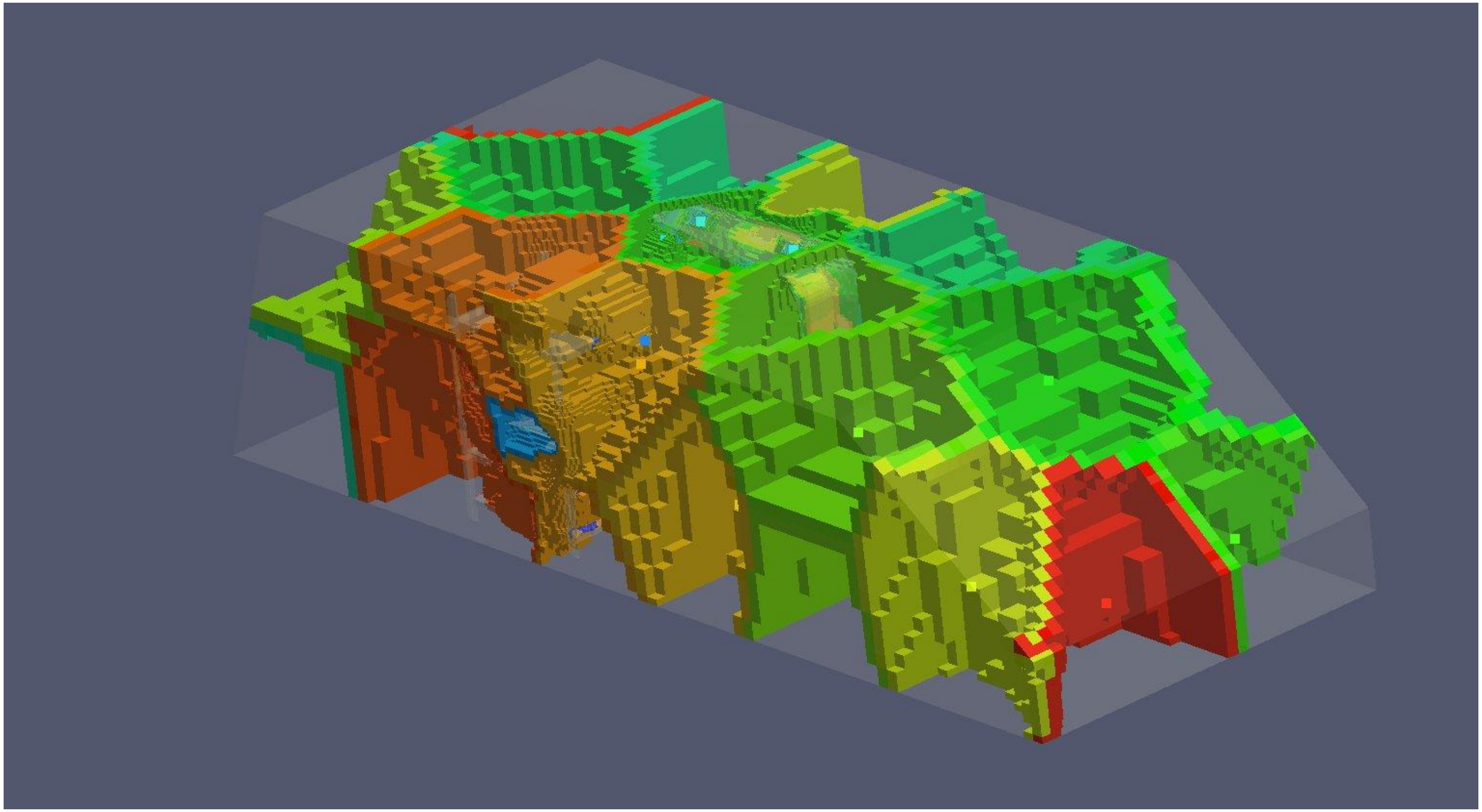
BACKUP SLIDES

50 Nodes



TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

50 Nodes

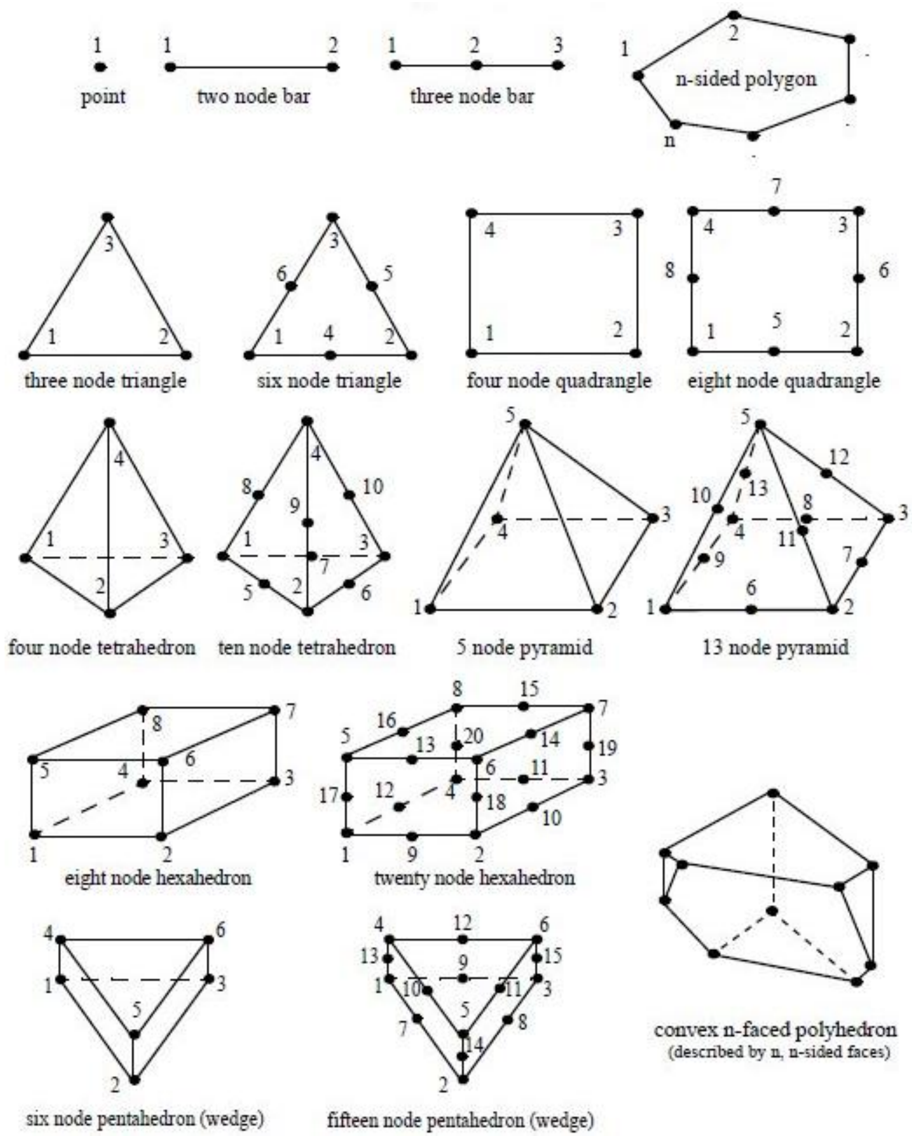


K-Means

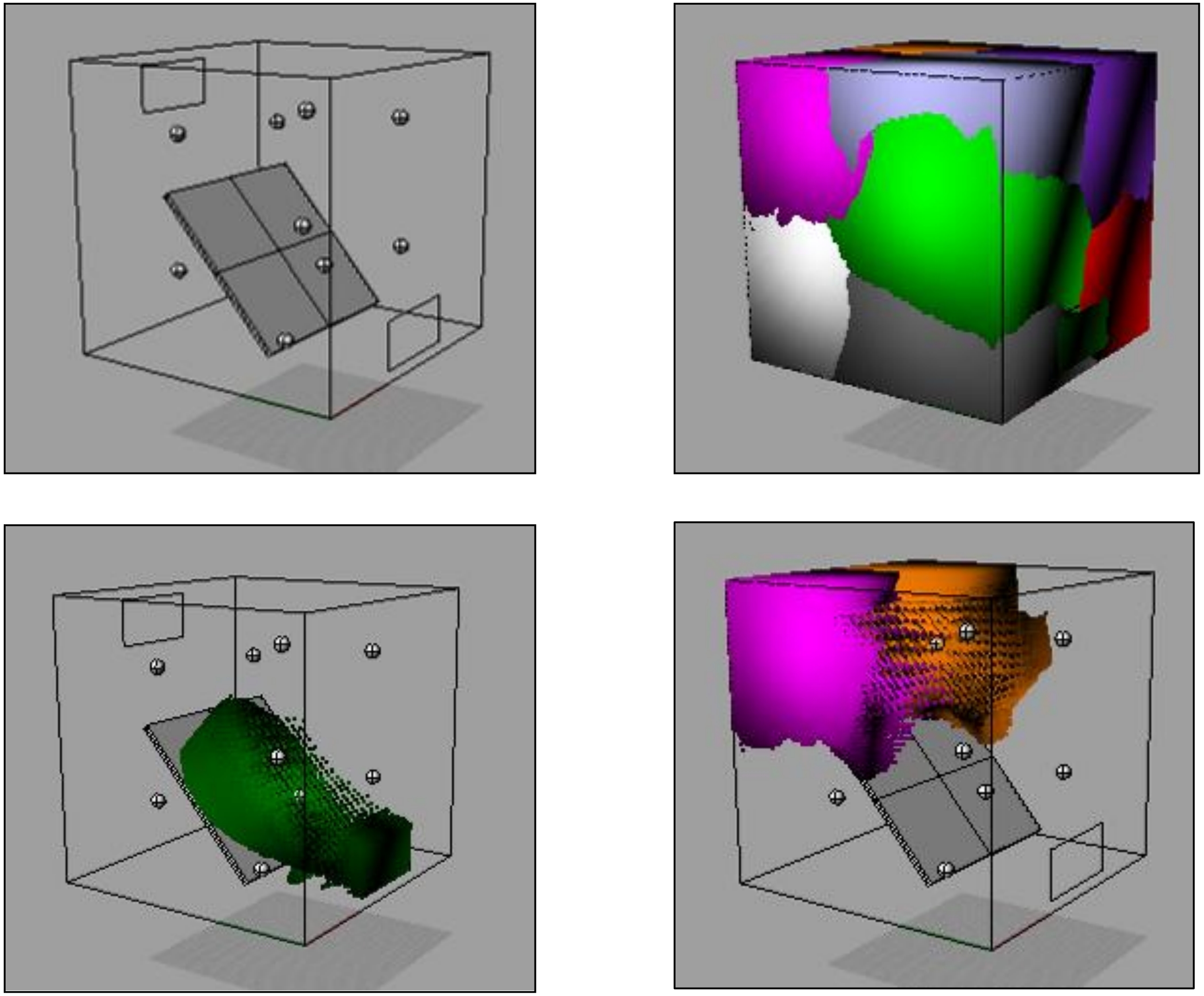
- K-Means was first suggested in 1962 by Sebestyen [15] and independently in 1967 by MacQueen [16]. A detailed history and full pedantic discussion of various deviations may be found in "K-Means Clustering: A Half-Century Synthesis" [17].
- Performance function: Minimize
  - $Sum\ Squared\ Error = \sum_{l=1}^{K} \sum_{x \in S_l} \left|\left|\vec{x} - m_l\right|\right|^2$
- K-Means Iterative Function
  - $\overline{m}_l = \frac{1}{n_l} \sum_{x \in S_l} \vec{x}$

K-Means Basic Algorithm

- for i=1 to number of tries
  - repeat
    - randomly select cluster nodes
    - compute distance of every point in domain to nodes
    - assign points to cluster based on minimum distance
    - compute volume weighted means to find new nodes
    - track the solution with minimum total error
  - until no change
- end for

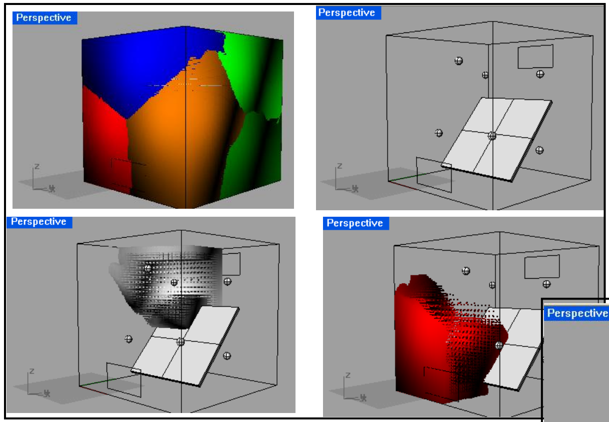

point
two node bar
three node bar
n-sided polygon
three node triangle
six node triangle
four node quadrangle
eight node quadrangle
four node tetrahedron
ten node tetrahedron
5 node pyramid
13 node pyramid
eight node hexahedron
twenty node hexahedron
six node pentahedron (wedge)
fifteen node pentahedron (wedge)
convex n-faced polyhedron
(described by n, n-sided faces)

Example Clustering

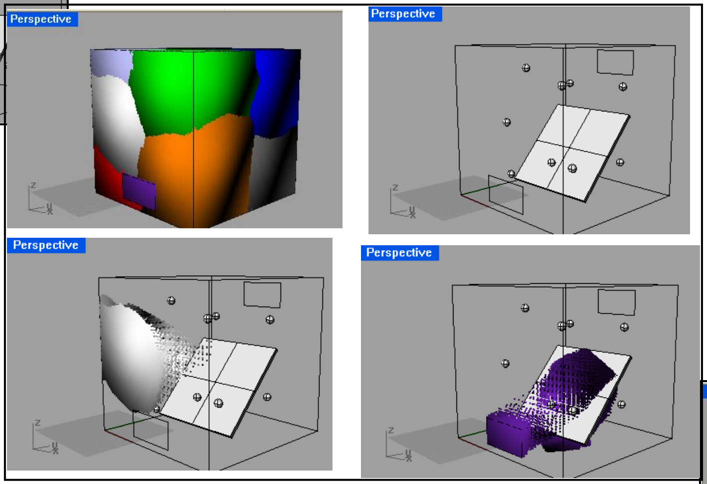


Number of Clusters

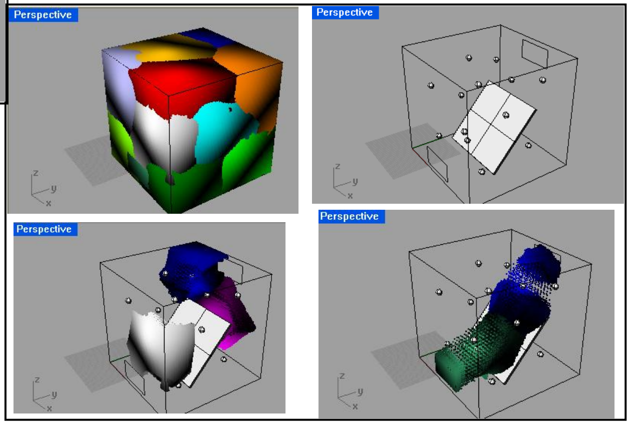
TARDEC



7 Clusters



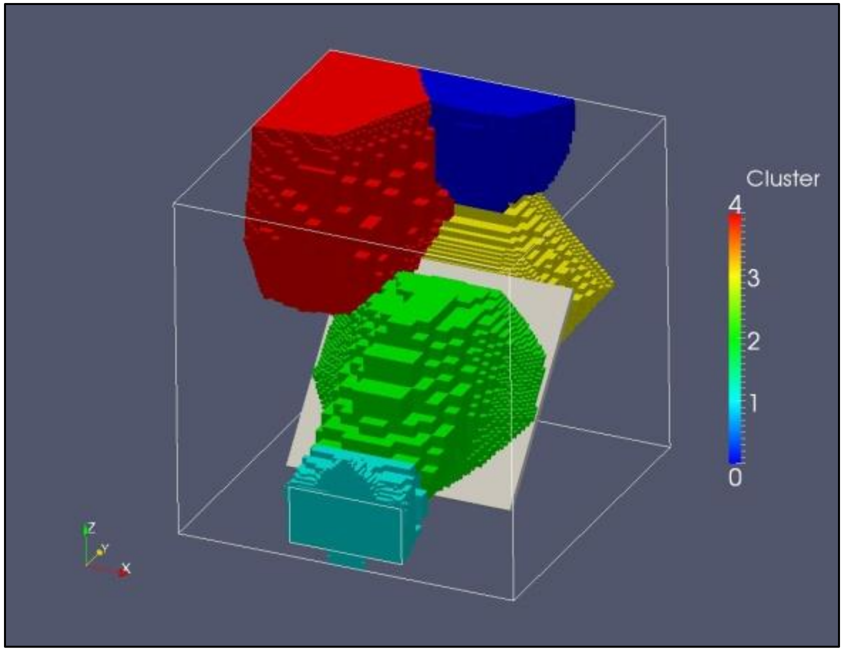
10 Clusters

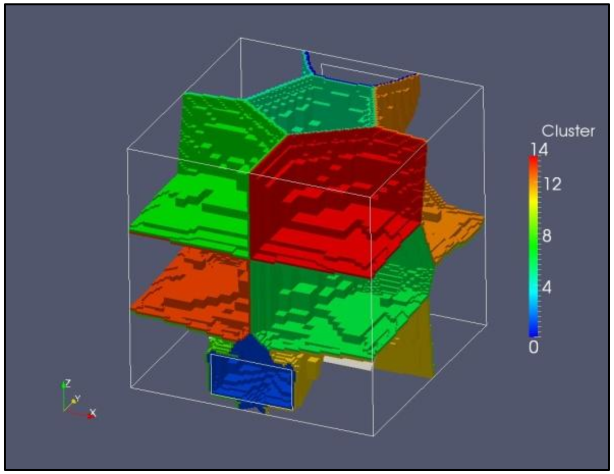


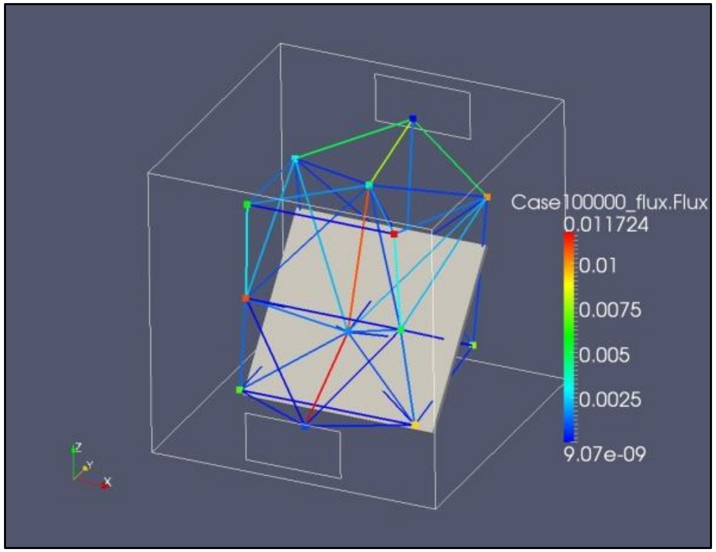
15 Clusters

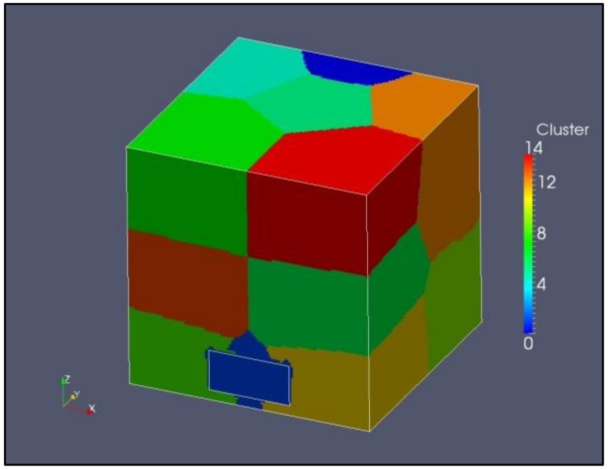
Result: 10 clusters is the minimum for this particular case to capture steam.

Visualization of Test Case 1

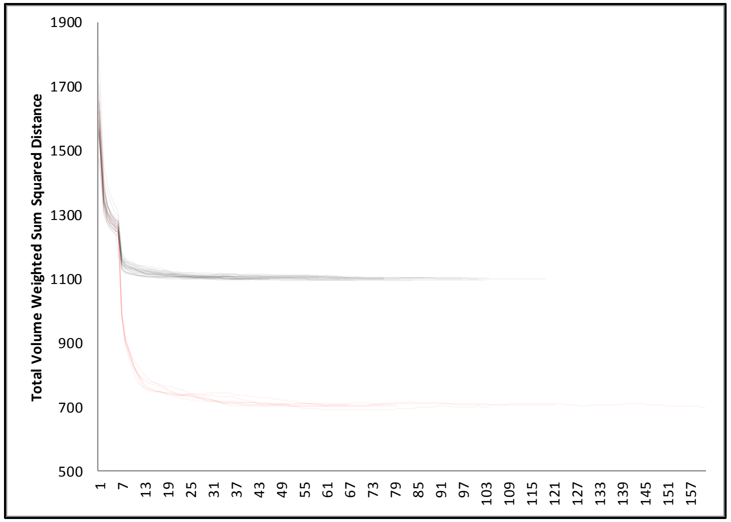


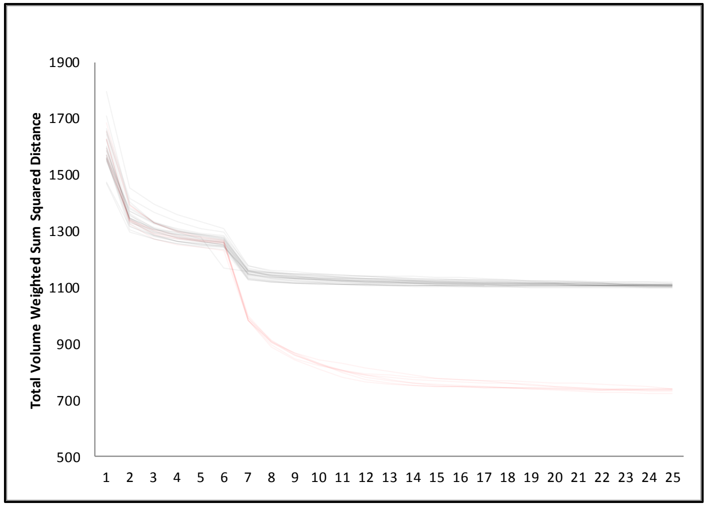






K-Means Mahalanobis Convergence





Typical k-means convergence plot for 33 random start trials on a vehicle duct geometry. The lines are translucent. The black lines show 25 runs with a 50% Mahalanobis/Euclidean blend versus the 8 red lines which show 75% Mahalanobis. Note the first 5 iterations are pure Euclidean

“Zoomed-in” plot of typical k-means convergence plot for 33 random start trials on an vehicle duct geometry. The lines are translucent. The black lines show 25 runs with a 50% Mahalanobis/Euclidean blend versus the 8 red lines which show 75% Mahalanobis. Note the first 5 iterations are pure Euclidean

Data Mining / Clustering

- Primary two categories for data clustering are:
    1. Hierarchical
    2. Partitional.
- Hierarchical clustering is based on the notion that objects are more related to nearby objects than to objects farther away.
    - Hierarchical clustering can be computed bottom-up or top-down. Hierarchical agglomerative clustering is starting with single elements and aggregating them into clusters.
    - Hierarchical divisive clustering is starting with the complete data set and dividing it into partitions.
- Partitional clustering operates on data globally
    - Exemplified by k-means where given an initial set of “k” means, points are assigned to the most similar “k” cluster.
    - Centroid-based clustering optimization is ultimately considered an np-hard problem, but is possible to solve for most practical cases.

New cfdMine Mahalanobis Algorithm

Mahalanobis cannot be used directly with K-means. Adds 36 unknowns and drops into non-optimal local minima even faster.

1. Pick random cluster centers from the domain and assign elements randomly
2. Z-normalize all the data in the domain

3. First m iterations are Euclidean
4. If iterations > n, compute the covariance matrix, otherwise identity matrix

5. Divide the x,y,z 3x3 inverse covariance matrix by the maximum value and multiply by the weight of importance for physical distance. For temperature, velocity magnitude, and velocity direction, divide the associated rows and columns by the diagonal value.
6. Blend the inverse covariance matrix with an identity matrix with variable weights on the diagonal.
7. Compute the Mahalanobis Distance from each domain point to the covariance matrix.
8. Assign each point the closest cluster.
9. Go to step 3 until converged